Ph.D. DISSERTATION

SPATIALLY RESOLVED CANCER GENOME ANALYSIS BY PHLI-SEQ (PHENOTYPE-BASED HIGH-THROUGHPUT LASER-AIDED ISOLATION AND SEQUENCING)

표현형 기반 고속 레이저 분리 및
염기서열 분석 기술을 이용한 암 유전체 분석

BY
SUNGSIK KIM

DECEMBER 2018

INTERDISCIPLINARY PROGRAM FOR BIOENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

SPATIALLY RESOLVED CANCER GENOME ANALYSIS BY PHLI-SEQ (PHENOTYPE-BASED HIGH-THROUGHPUT LASER-AIDED ISOLATION AND SEQUENCING)

지도 교수 권 성 훈

이 논문을 공학박사 학위논문으로 제출함
2019 년 2 월

서울대학교 대학원 협동과정 바이오엔지니어링
김 성 식

김성식의 공학박사 학위논문을 인준함
2019 년 2 월

위 원 장 한 원 식 (인)

부위원장 권 성 훈 (인)

위 원 박 웅 양 (인)

위 원 백 성 희 (인)

위 원 방 두 희 (인)

Abstract

Despite advances in next-generation sequencing technologies, it has been challenging to map molecular information of cells to their tissue context due to technical limitations. In this dissertation, I describe PHLI-seq, a novel approach that enables high-throughput isolation and genome-wide sequence analysis of single-cells or small numbers of cells from a tissue to construct genomic maps within biological samples in relation to the images or phenotypes of the cells. By applying PHLI-seq to breast cancer tissues, genetic heterogeneity of cancer is revealed at a high spatial resolution. Also, in a case of ovarian cancer, a history of tumor development is explored using PHLI-seq. Through these studies, it is concluded that genetic subclonality and spatial separation of the subclones are formed in early stage of cancer development.

Keywords: Genomics, Cancer, Next-generation sequencing, Spatially resolved sequencing, Laser microdissection

Student Number : 2013-30999

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advances in sequencing technology have revolutionized biological and medical researches. This revolution provided an opportunity to overcome the limitations of conventional analysis techniques for biology [1]. Despite the remarkable advances in sequencing technology, the foundations of modern oncology lie in histopathology, which remains the gold standard for comprehending tumorigenesis, relapse, metastasis, cancer evolution, and appropriate clinical applications [2]. However, the conventional histopathological analyses often fail to identify detailed molecular information in whole-genome or whole-transcriptome. This loss of information for linking genomic information with the histopathological context hinders deeper analyses of the tumor mass. Thus, it is important to link the spatial and phenotypic histopathological information to the objective and massively parallel measurement of the molecular alteration status of cells in a high-throughput manner for precise understanding of cancer and improvement in diagnostic performance [3]. Therefore, spatially resolved

genomics has emerged to address this issue. Considering recent progress in spatially resolved transcriptomics [19−22], technical advances are required to map the genomic data for cancer spatially.

1.1 Identifying Molecular Information of Cells

Cells are the basic unit of living creatures. Thus, fundamentals and mechanisms of biology can be explored by studying cells. There are various approaches for studying cells. For example, analysis of biochemical composition of cells can provide information about the status of cells. Also, studying metabolic and signaling pathways of cells can provide insight on the mechanism that controls and coordinates the actions of cells. Lastly, research on interactions of the cells with other cells or their environment can teach us how cells work together as a part of an organism.

Many pathological examination and medical diagnostics also rely on analyzing cells. Medical experts or medical devices observe or measure any abnormalities in samples from patients. For test, patients usually provide their blood or tissues. Then, the samples are analyzed by various methods such as microscopy, cytometry, molecular staining, or nucleic acid sequencing. In these analysis methods, cells from patients provide valuable information about the status of the patients.

Among various methodologies for analyzing cells, molecular information of cells provides broad range of knowledge about biology and medicine. Molecular information of cells can be collected from DNA, RNA, or protein of cells. Recently, with full advantages of high-throughput techniques, measuring whole genome, transcriptome, and proteome have been possible to open the era of data-driven biology and medicine.

For getting molecular information of cells, conventional and most commonly used method is molecular staining. The basic principle of molecular staining is attaching a visible marker to a target molecule of cells. A visible marker can be dyes, metals, or fluorescent molecules. A target molecule of cells can be DNA (usually certain gene sequence), RNA, or protein (usually surface proteins). For example, immunohistochemistry (IHC) method stains cells with a certain protein. This method uses antigen-antibody interaction to target a specific protein. Similarly, immunofluorescence (IF) method uses antibody to stain cells with a certain protein, but imaging after staining is performed with fluorescent microscopy. Whereas, fluorescence *in situ* hybridization (FISH) uses short sequence of DNA or RNA as a probe for targeting a certain sequence of DNA or RNA of cells. In this method, a fluorescence dye is attached to a probe sequence, and specificity of targeting comes from the probe sequence.

The molecular staining methods are useful to identify information of a few molecular targets. However, a limited number of target molecules restricts the usefulness of the staining methods. In contrast, next-generation sequencing (NGS) technology can fulfill the needs of analyzing whole genome, transcriptome, or proteome. NGS technology is now utilized in biological subjects and medical diagnosis such as development [8], stem-cell research [9], cell type identification [10], pathogenic gene study [11], and cancer research [12].

After development of NGS technology, researchers have invented many technical applications of the technology. One of the high-impact applications is single-cell sequencing technique. Single-cell sequencing examines the sequence

information from individual cells with optimized NGS technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell [13]. Single-cell sequencing technique requires cells to be dissociated into single-cells for preparing individual sequencing libraries. In addition, dissociated single-cells must undergo lysis and sequencing library preparation individually.

There are several different approaches for single-cell sequencing. In the past, researchers used micromanipulators or mouth pipette for isolating single-cells [14]. More recently, fluorescence activated cell sorting (FACS) has been used to isolate single-cells into different reaction tubes [15]. However, the two methods have their limitation in parallelized sequencing library preparation. In other words, researchers should prepare a sequencing library for each cell, even if single-cells are separated by a high-throughput technique such as FACS. Thus, techniques have been emerged that enable parallelized single-cell separation and sequencing library preparation.

Two representative techniques of parallelized single-cell sequencing techniques are droplet-based [16]−[18] and microfluidics-based methods [6]. In principle, both methods have similar aspects in that a single-cell is confined to an isolated container where independent lysis and sequencing library preparations take place. Most importantly, each container has a unique molecular identifier which gives a unique DNA sequence tag at the end of sequencing library of each cell. Therefore, researchers have become able to sequence hundreds to thousands of single-cells through a simple experiment.

1.2 Spatially Resolved Sequencing of Cells

In the previous section, techniques for identifying molecular information of cells have been described. However, for deeper understanding of the biology of multicellular organisms, we must devise ways to perform high-throughput measurements of molecular information while preserving spatial information about the tissue context [19]. In this section, I will describe techniques that can provide both spatial information of cells and the cells' sequence information produced by high-throughput sequencing technologies.

In situ sequencing [20], [21] and single-molecule fluorescence *in situ* hybridization (smFISH) [22] are typical examples of the spatially resolved sequencing techniques. *In situ* sequencing techniques use sequencing probes for targeting cDNA in targeted cells. Then, sequencing-by-ligation chemistry generates sequential fluorescent signals from the cells. By analyzing the fluorescent signals, researchers are able to uncover gene expression profiles of the cells. In contrast, fluorescent labeled probes are directly utilized in the smFISH techniques. The probes have unique sequences for targeting specific cDNAs, and gene expression profile is analyzed by reading the fluorescent signal of the probes. Since both *in situ* sequencing and smFISH are solid-phase experiments, or carried out on the surface of slide glasses, biological samples can be analyzed without dissociation. However, a critical disadvantage of these techniques is the inability to use the high-throughput sequencing capacity of NGS, and researchers analyze a few genes at a time with these techniques. Therefore, *in situ* sequencing and smFISH are used only in limited circumstances, although these techniques have a

great potential to be utilized for studying spatially resolved gene expression profiles.

Junker *et al.* also suggested an approach, called tomo-seq, which can provide spatial and sequence information at the same time [5]. Furthermore, tomo-seq uses NGS to generate sequence information. In tomo-seq, by sectioning individual samples in three different directions (ex. x, y, and z direction), RNA profiles along the three main body axes are sequenced. Then, with the algorithm inspired by computed tomography (CT) technology, spatial gene-expression patterns in 3D are reconstructed on a transcriptome-wide level.

Similar to tomo-seq, Laser Microdissection (LMD) technique is used for sectioned biological samples. However, unlike tomo-seq, there is a difference in that sequencing is done after separating a desired part in one section. Since LMD can be performed with histological imaging, accurate connections between spatial and sequence information can be achieved. Usually, more than thousands of cells are separated into an experimental tube, though some researchers used LMD for isolating and sequencing in 1-cell resolution [23], [24].

1.3 Comparison of Technologies

As we discussed in the previous sections, it is important to identify molecular and spatial information at the same time for better understanding of biology (Figure 1.1). There are various methods for this purpose, but all method has its strength and weakness. For example, droplet-based single-cell sequencing technique can produce over 10,000 single-cell sequencing libraries through one experiment, but

this method requires cells to be dissociated before experiment. Therefore, this technique loses spatial information of the analyzed cells. On the other hands, *in situ* sequencing can preserve spatial and histologic information by nature. However, this is neither genome-wide nor single-cell resolution sequencing method, which means *in situ* sequencing has its weakness in identifying molecular information of cells.

Likewise, all methods have advantages and disadvantages, and several representative techniques are presented in Figure 1.2 according to three aspects: 'cell identity', 'histology' and 'throughput'. Here, you can notice that techniques such as single-cell sequencing, FACS, and conventional NGS lose spatial or histologic information, since cells must be dissociated or lysed before applying these methods. Whereas, *in situ* sequencing, smFISH, and histologic imaging are basically measuring molecular information by microscopy, and these methods cannot take advantage of NGS. LMD can be used to bridge the gap histologic imaging with NGS, but it has two critical disadvantages: throughput is low, and the

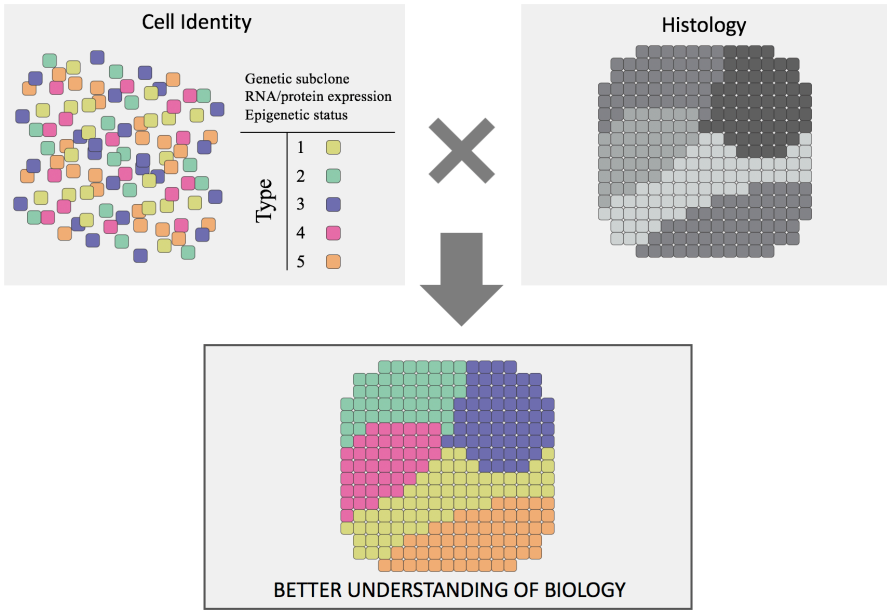

Figure 1.1 For better understanding of biology, it is required to consider cell identity and histology at the same time

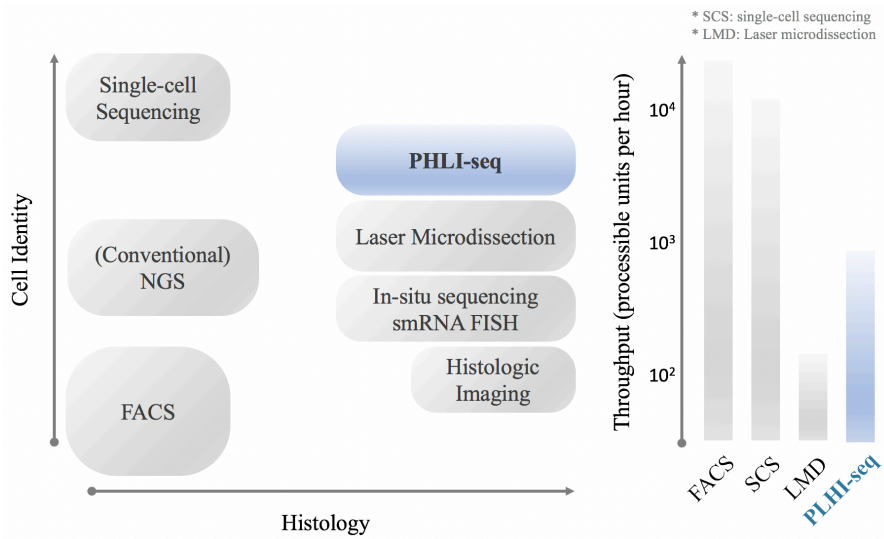

Figure 1.2 Technical positions of various methods for identifying molecular information of cells

method usually applied to isolate and sequence thousands of cells at once. Therefore, there has been a room for a technique which can analyze a small number of cells with NGS and preserve histologic information. With the technical needs, phenotype-based high-throughput laser-aided isolation and sequencing (PHLI-seq) was developed.

Chapter 2

Technical Overview of PHLI-seq

In this chapter, phenotype-based high-throughput laser-aided isolation and sequencing (PHLI-seq) is described [25]. PHLI-seq can efficiently detect tumor subclonality and variants with low-level allele fractions with high sensitivity and accuracy while preserving information of spatial organization and histopathological images of cells. For PHLI-seq analysis, cells are physically isolated into different tubes by discharging cells on an indium tin oxide (ITO)-coated glass slide using an infrared laser pulse. Finally, the extracted DNA from each tube is amplified and sequenced. The procedure is automated with custom software to achieve a high-throughput analysis. By establishing this novel approach, researchers can identify distinct subclones and their somatic variants in various samples.

2.1 Phenotype-based High-throughput Laser-aided Isolation and Sequencing (PHLI-seq)

For isolating cells in PHLI-seq technique, a single shot of focused near-infrared laser (1064 nm) was applied. As the infrared laser shot vaporized ITO, the prepared cells on the targeted area were transferred downward by the pressure (Figure 2.1a). In contrast to laser microdissection (LMD) using UV cutting and catapulting [9, 25] , PHLI-seq which uses an infrared laser and one-shot isolation step causes no damage to the nucleic acids and can be performed rapidly, enabling to recover high-quality DNA from single cells or small numbers of cells in high-throughput. On the other hand, standard LMD using an infrared laser to heat a thermoplastic film expands and captures the cells of interest. LMD which uses infrared laser is free of UV-induced damage, but the thermoplastic film should physically contact with samples, and tearing off the cells by removing the cell-captured cap. This physical contact causes inherent problems, such as limited throughput and scalability; debris or non-selected cell adhesion; and limited applicability depending on the sample condition [10, 11]. Through PHLI-seq, we isolated regions of cells or even single cells from various samples, regardless of how closely positioned the cells were, while preserving information regarding the spatial and histopathological information (Figure 2.1b). Together with the operating software, it was able to use PHLI-seq to isolate targets (1 second per target) into retrieval tubes in high-throughput.

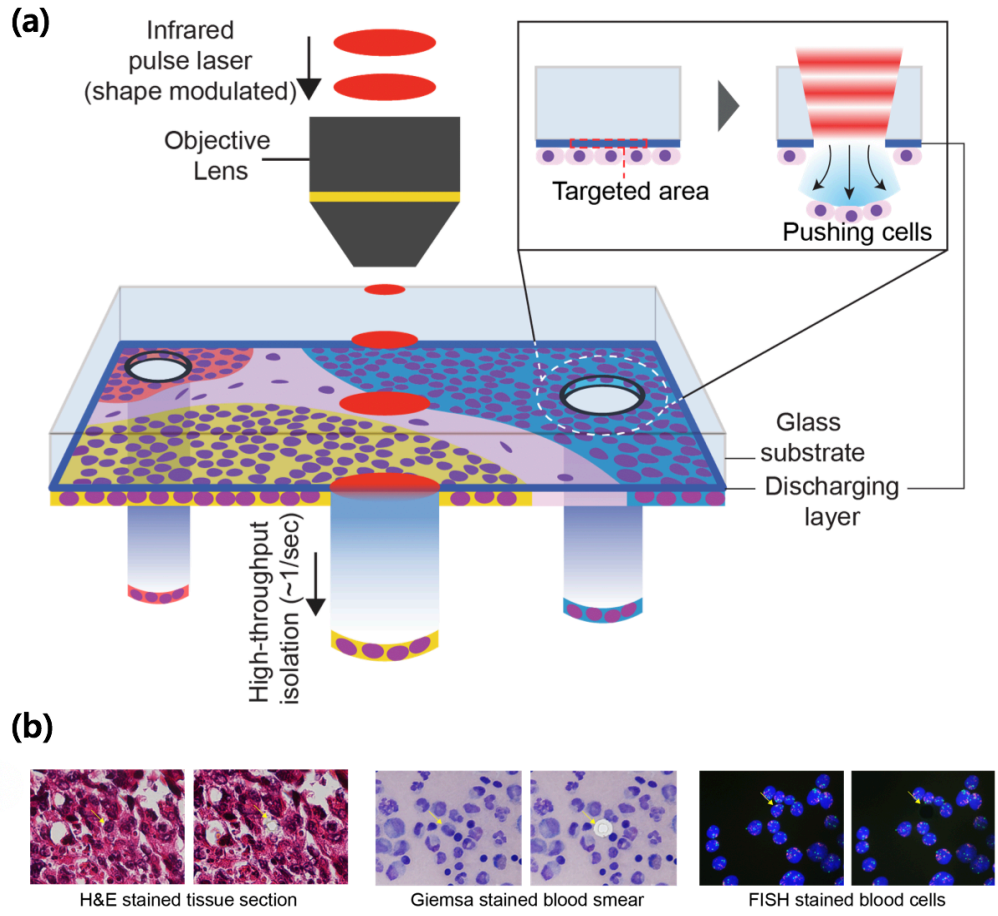

Figure 2.1 Phenotype-based High-throughput Laser-aided Isolation and Sequencing (PHLI-seq) (a) PHLI-seq uses infrared laser pulses and discharging layer (ITO) coated glass slides to isolate cells. (b) PHLI-seq can be applied to various sample types [25].

ITO is a transparent material in the visible light range. However, absorption spectrum of ITO shows that the material absorbs larger amount of light in the UV and IR than visible region (Figure 2.2). More quantitatively, a biological tissue would absorb about 1,000-fold less energy than ITO at 1064 nm wavelength ($0.04/e^{3.7} \sim 0.001$). Moreover, biological tissues have about 10-fold higher heat capacity than ITO [30]. Since vaporization temperature is ~3000K for ITO, the layer can be vaporized without heat damage on biological materials. Experimentally, 100 nm of ITO layer could be vaporized by ~4 J/cm^2 of IR laser pulse (Figure 2.3).

2.2 Instrumentation

The PHLI-seq instrument comprises two motorized stages, a CCD camera, light source, laser source, pulse slit, objective lenses, and fluorescence modules **(**Figure 2.4), and all components are automatically controlled by a computer. One stage is for loading sample slides, and the other is for loading tubes to receive isolated cells. The CCD camera (Jenoptik, Jena, Germany) works for imaging samples through the objective lenses, and an Nd:YAG nanosecond laser was purchased from Continuum (Minilite^{TM} Series ML II; Continuum, San Jose, CA). A slit is located in the light path between the laser source and the objective lens to control the region to be isolated. Objective lenses from 4x to 50x magnifications were purchased from Mitutoyo.

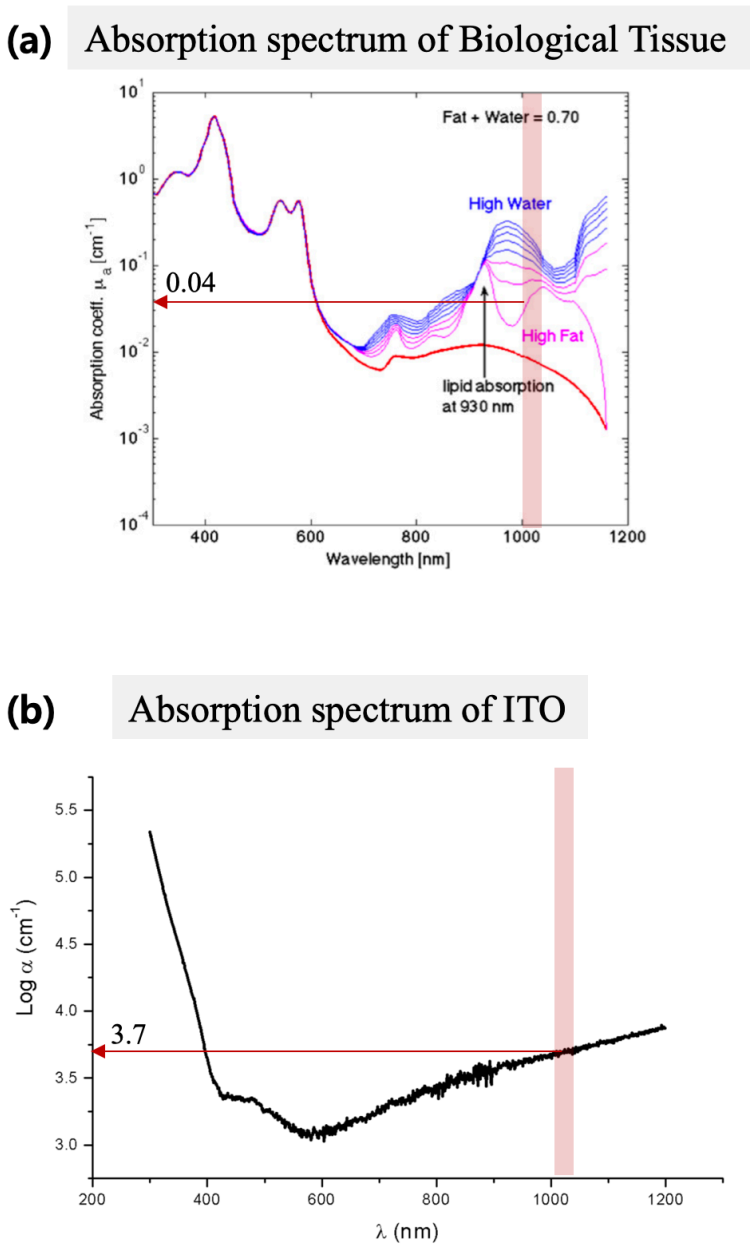

Figure 2.2 Light absorption spectrum of biological samples and ITO [31][32]

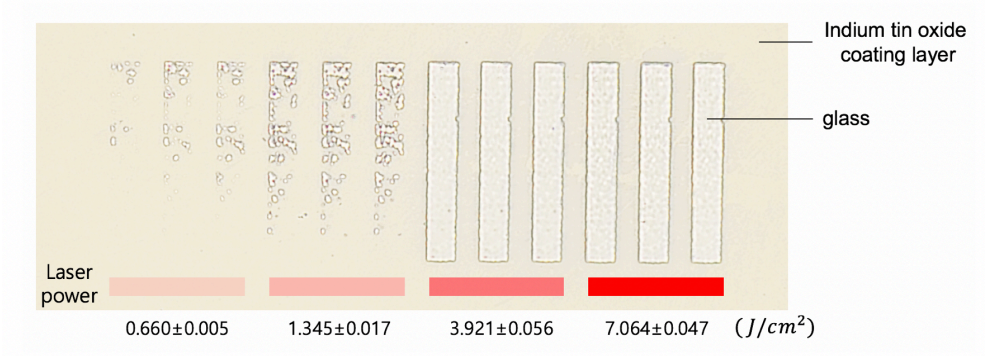

Figure 2.3 Vaporization of ITO coating according to various laser irradiance. 100nm thickness of ITO laser can be vaporized under ~4J/cm2. The laser powers were measured using laser power meter at the position where cells were located [25].

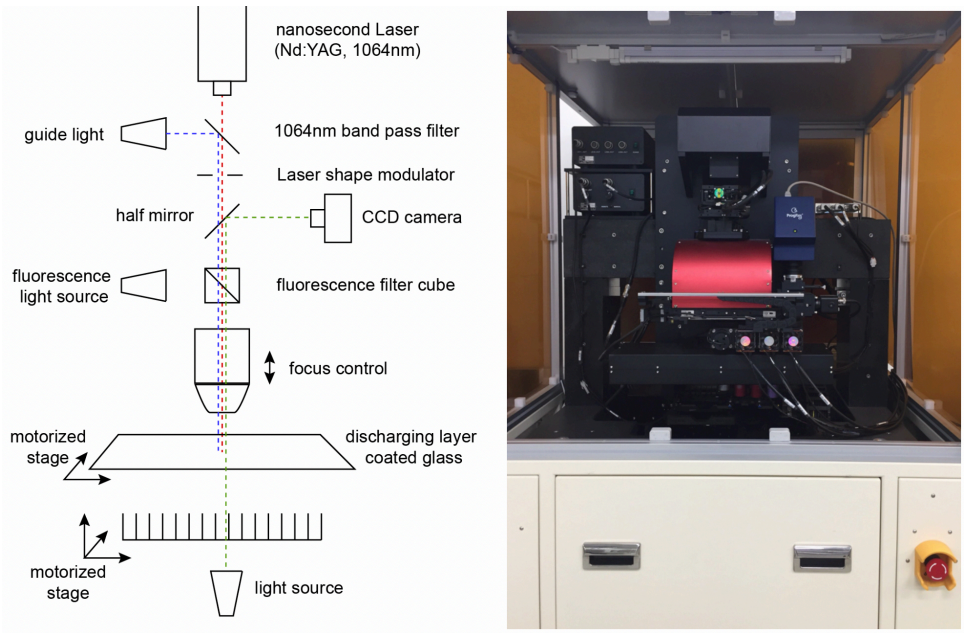

Figure 2.4 The schematic and the image of the instrument for PHLI-seq. The PHLI-seq instrument is comprised of two motorized stages, CCD camera, light source, laser source, laser shape modulating slit, objective lenses, and fluorescence modules. The entire system is controlled by in-house control software [25].

Two different pieces of software were created, which were written in Python scripts, and the codes are available at Github and Zenodo:

https://github.com/BiNEL-SNU/PHLI-seq

http://doi.org/10.5281/zenodo.1342126

One software was built for users (for example, a pathologist in a hospital) to select cells to be isolated. Whole slide images could be shared with users through a server, and the users ran the software to select cells of interest while navigating the tissue images through the graphical user interface. After running the software, two files are created. One is a text file with locational information about the region of interest, and the other is an image file with the selected targets overlaid with marks on the original image. Both files are required for the automated isolation of the target cells. The second software makes it possible to control the PHLI-seq instrument.

For isolating cells using PHLI-seq instrument, a retrieving cap holder was used as described by the Figure 2.5. This is a custom device to hold twelve 8-strip caps, which enable to capture up to 96 samples in a single run. The retrieving cap holder was designed for convenience, and it's not a prerequisite. The retrieving cap holder comprises a rack structure where twelve 8-strip caps can be mounted and beam structures which press down the 8-strip caps. The beam structures are attracted toward the rack structure by magnetic force in order to hold caps tightly.

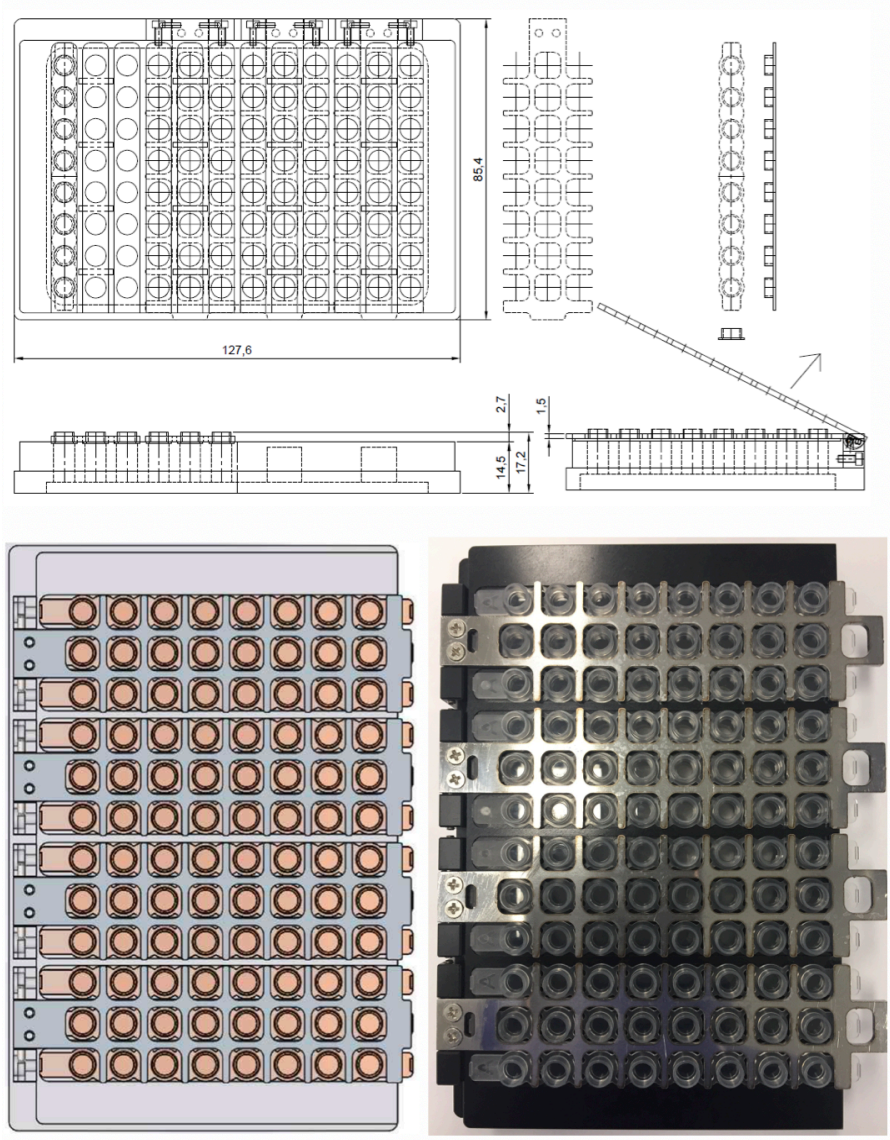

Figure 2.5 The design and the image of a retrieving cap holder for PHLI-seq [25]

Figure 2.6 shows the situation where the above retrieving cap holder is mounted on the PHLI-seq instrument to isolate cells. The slide glass holder is about 1.3 mm in thickness, and sample-to-cap distance is about 1.4~1.5 mm in actual operation. The retrieving cap has an inner diameter of 4.5 mm. Thus, when a cell is separated at an angle of under 60 degrees, it is accurately inserted into the cap. When a cell is separated at an angle of 72 degrees, the cell could enter into another cap.

Cells can be isolated from tissue sections, cell lines, or blood smears that were spread on the ITO glass, where ITO was coated on a glass by sputter deposition. Glass slides with 100-nm-thick ITO layer were used. The thickness of the ITO layer could affect the maximum tissue thickness at which the cells can be detached, although a thorough assessment was not undertaken. Using an infrared laser in conjunction with the ITO discharging layer, it was possible to isolate cells without damaging them (Figure 2.7).

2.3 Whole Genome Amplification

After isolating cells, PCR tubes were centrifuged, and cells were lysed using proteinase K (cat no. P4850-1ML, Sigma Aldrich) according to the manufacturer's directions. When a small number of cells were isolated into a tube, the total amount of nucleic acid would be not enough for post-processing. In this case whole-genome amplification was performed to amplify the amount of nucleic acids. There are several methods for whole-genome amplification. Here, DOP-PCR, MDA, and MALBAC will be described.

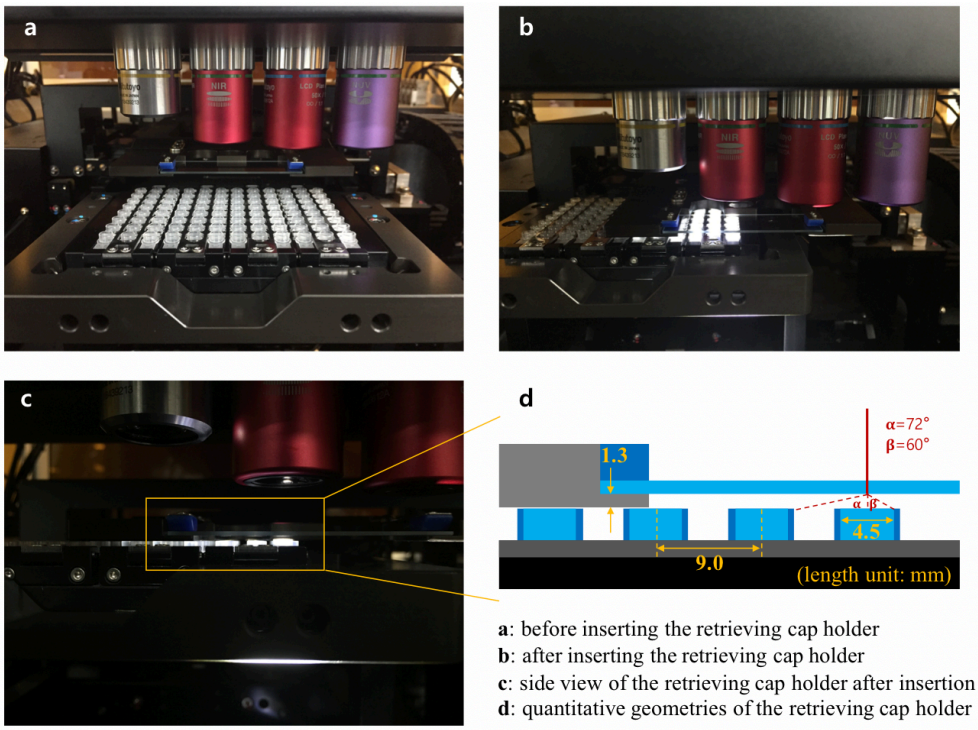

Figure 2.6 The images show situations where the retrieving cap holder is mounted on the PHLI-seq equipment to isolate cells [25].

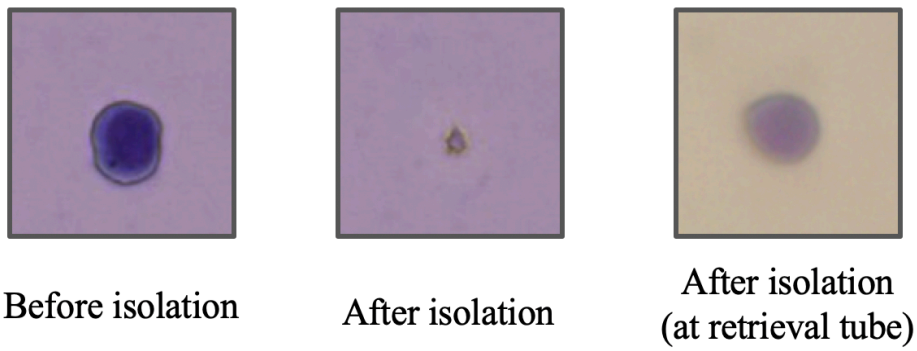

Figure 2.7 Isolation of a cell line through PHLI-seq. The images of before and after isolation show that the cell was clearly isolated from a slide glass. The scar in the "After isolation" is the region of vaporized ITO. At the retrieval tube, the isolated cell could be observed without damages [25]

DOP-PCR (degenerate-oligonucleotide-primed PCR) is a type of whole-genome amplification method which uses degenerate-oligonucleotide as a primer for amplification. DOP-PCR consists of two different cycling stages. The first is the low-temperature annealing stage where DOP primers anneal and tag genomic sequences. In the second stage, the PCR annealing temperature is raised, which increases priming specificity during amplification of the tagged sequence [33].

MDA (multiple displacement amplification) is another type of whole-genome amplification method. In contrast to DOP-PCR, MDA is an isothermal whole genome amplification method. The reaction starts by annealing random hexamer primers to the template at 30°C. Then, DNA synthesis is carried out by Φ29 DNA polymerase. Since Φ29 DNA polymerase has high processivity and strand displacement activity, MDA generates larger sized products (10kb ~ 100kb).

MALBAC (Multiple Annealing and Looping Based Amplification Cycles) is a recently developed method for whole genome amplification. This method also comprises two steps. The first step is tagging genomic sequences to MALBAC primers using high-processivity polymerase. The second step is amplifying the tagged sequence using thermocycling. Therefore, this method can be interpreted as a combination of DOP-PCR and MDA. However, MALBAC utilizes special primers that allow amplicons to have complementary ends and therefore to loop. This looping prevents amplified DNA from being amplified again, which can reduce amplification bias of genomic sequences. [34].

Each method has its advantages and disadvantages. For comparing the three methods, whole genome amplification can be evaluated in the three point of views (Figure 2.8): coverage-uniformity, coverage-breadth, and false-positive rate. Ideally, high coverage-uniformity and coverage-breadth and low false-positive rate is preferred, but the three whole genome amplification methods have their own weaknesses. For PHLI-seq in the following experiments, MDA was chosen for whole-genome amplification, because high coverage-breadth and low false-positive rate were important for following applications.

For whole-genome amplification by multiple displacement amplification, GE's Illustra Genomiphi V2 DNA amplification kit (cat no. 25-6600-30) was used. The manufacturer's instruction was followed, except that 0.2 μl of SYBR green I (Life Technologies) was added into the reaction solution for real-time monitoring of the amplification. All amplified products were purified using Beckman Coulter's Agencourt AMPure XP kit (cat no. A63880) immediately following the amplification reaction. After amplification, it is required to filter out badly

amplified products. For this, we used real-time whole-genome amplification monitoring and PCR validation with in-house designed 16-region primer panels (Table 1). To discard poorly amplified samples, I used MDA products that exhibited an observable amplification level within 40 minutes after the start of the reaction. Most of the amplified products yielded more than 1 μg, among which 800 ng was used for Illumina library construction. To prevent carry-over contamination, the pipette tip, PCR tube, and cap for the reaction were stored in a clean bench equipped with UV light and treated with O_2 plasma for 30 seconds before use. Moreover, non-template controls were included in every experiment for ensuring that no contaminants were transferred.

Method	Uniformity of coverage	Coverage breadth	False-positive
MDA	Low	High	Low
MALBAC	Meidum	Medium	High
DOP-PCR	High	Low	Low

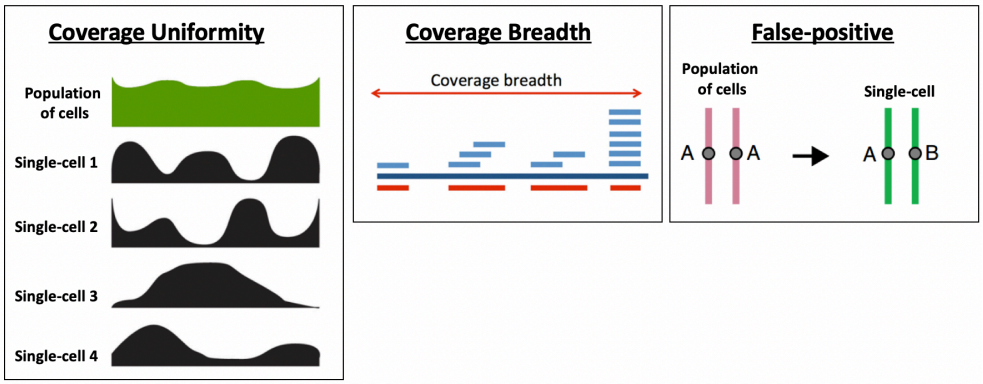

Figure 2.8 Comparison of three whole genome amplification methods: MDA, MALBAC, and DOP-PCR. Coverage uniformity, coverage breadth, and false-positive rate are the key observables for evaluating whole genome amplification [35]

Table 1 Primer sets for validating whole-genome amplified products

Validation primer 1	F	5'-TCTAGACCTGCCACTGGGAA-3'
	R	5'-ATGCAGCAGGTGCTGAGTAA-3'
Validation primer 2	F	5'-ACTGCCCATGCACTTTGACT-3'
	R	5'-CCACACTCCTTCGCCAACTT-3'
Validation primer 3	F	5'-ACACCATGAAGCAGAAGGGG-3'
	R	5'-TGCATGAGCCCATGTACCTC-3'
Validation primer 4	F	5'-GGATGACTGGAGCAGGGAAG-3'
	R	5'-TGGGCAGCATCCATTGAGAG-3'
Validation primer 5	F	5'-AAGAGCATTTTTATGCTCCATCTG-3'
	R	5'-CACATACAGACCCGCTGGAA-3'
Validation primer 6	F	5'-GCCAACATGGCCAGGAAGTA-3'
	R	5'-TCATGTGCACAAATGTATGTTTCTT-3'
Validation primer 7	F	5'-GGAAGGCTTTGAAGAAGGTGAAT-3'
	R	5'-AGCACCAAAAAGGCACATACC-3'
Validation primer 8	F	5'-AACCCTCCCAATTCCAGTGC-3'
	R	5'-ACAGTTCTTTTCATCACTGCCG-3'
Validation primer 9	F	5'-GAGCCACATGAGTCTGCCAT-3'
	R	5'-AGAGCCAGGCTTTTGCTGAA-3'
Validation primer 10	F	5'-CTTCCTTGGGGACCACATCC-3'
	R	5'-CCCATCGTCTCTGCTGACAA-3'
Validation primer 11	F	5'-GTGTGCGGAAGGTACGGTTA-3'
	R	5'-TTGCTCCTGCTCAGGTCTTG-3'
Validation primer 12	F	5'-TCAATCTCCATGCCCAGGGT-3'

	R	5'-TTCACTGCCAACATTGCACG-3'
Validation primer 13	F	5'-GTGGACAGCTGACACGAGAG-3'
	R	5'-CGAGAGGCCACAGAAGTAGC-3'
Validation primer 14	F	5'-AGGTACCCGTACATACCAGGA-3'
	R	5'-TGGTGCTGGCAGGATAACAG-3'
Validation primer 15	F	5'-CTTGCTGGTCTGTCCCTCTG-3'
	R	5'-ATCCTCCCCCACCTCCTTTT-3'
Validation primer 16	F	5'-ACCTCAGCAACCTTCAAGAACT-3'
	R	5'-GGGTGTAGAATCAAACCAGCG-3'

2.4 Performance Validation of PHLI-seq

To calculate the collection efficiency and the probability of obtaining high-quality sequencing data after cell isolation, single-cells or very small numbers of cells were isolated and sequenced from a cell line or an H&E-stained fresh frozen cancer tissue section. For calculating collection efficiency, HL60 cells were isolated and whole genome amplified. Amplification procedure was monitored using fluorescent dye and a threshold was set to determine successfully performed trials. Through this experiment, it was determined that the collection efficiency is 92.5% (49/53) from single-cell isolations of the cell line sample (Figure 2.9). For the experiments with H&E-stained tissue sections, nucleic acids in the cells could be damaged during preservation, tissue sectioning, and staining. The damages can affect whole-genome amplification in such a way that coverage uniformity is reduced, and the false-positive rate is increased. Therefore, the probability of obtaining high-quality sequencing data was calculated from 5-cell and 1-cell isolations from H&E-stained tissue sections. It was found that 81.3% (13/16) and 18.8% (3/16) of the 5-cell and 1-cell isolation experiments produced high genomic coverage and correct copy number profiles (Figure 2.9). Based on the sequencing data, it was also found that high genomic coverage and a high correlation with the true copy number value of the whole-genome amplified samples were related to low C_T values in real-time monitoring of the amplification reactions (Figure 2.10). Therefore, for following experiments, it was decided to monitor the MDA procedure using a real-time PCR machine and validated the products by PCR using

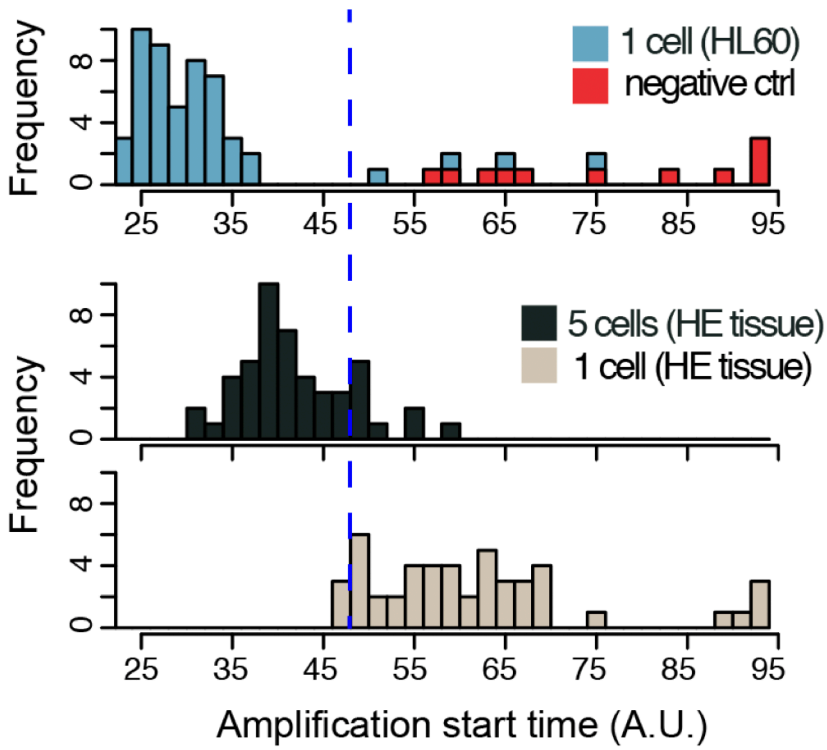

Figure 2.9 Real-time monitoring of whole genome amplification of the HL60 cell line and H&E stained cells. Amplification start time was measured for each isolation for calculating collection efficiency and the probability of obtaining high-quality sequencing data. The blue dashed line is the threshold value [25].

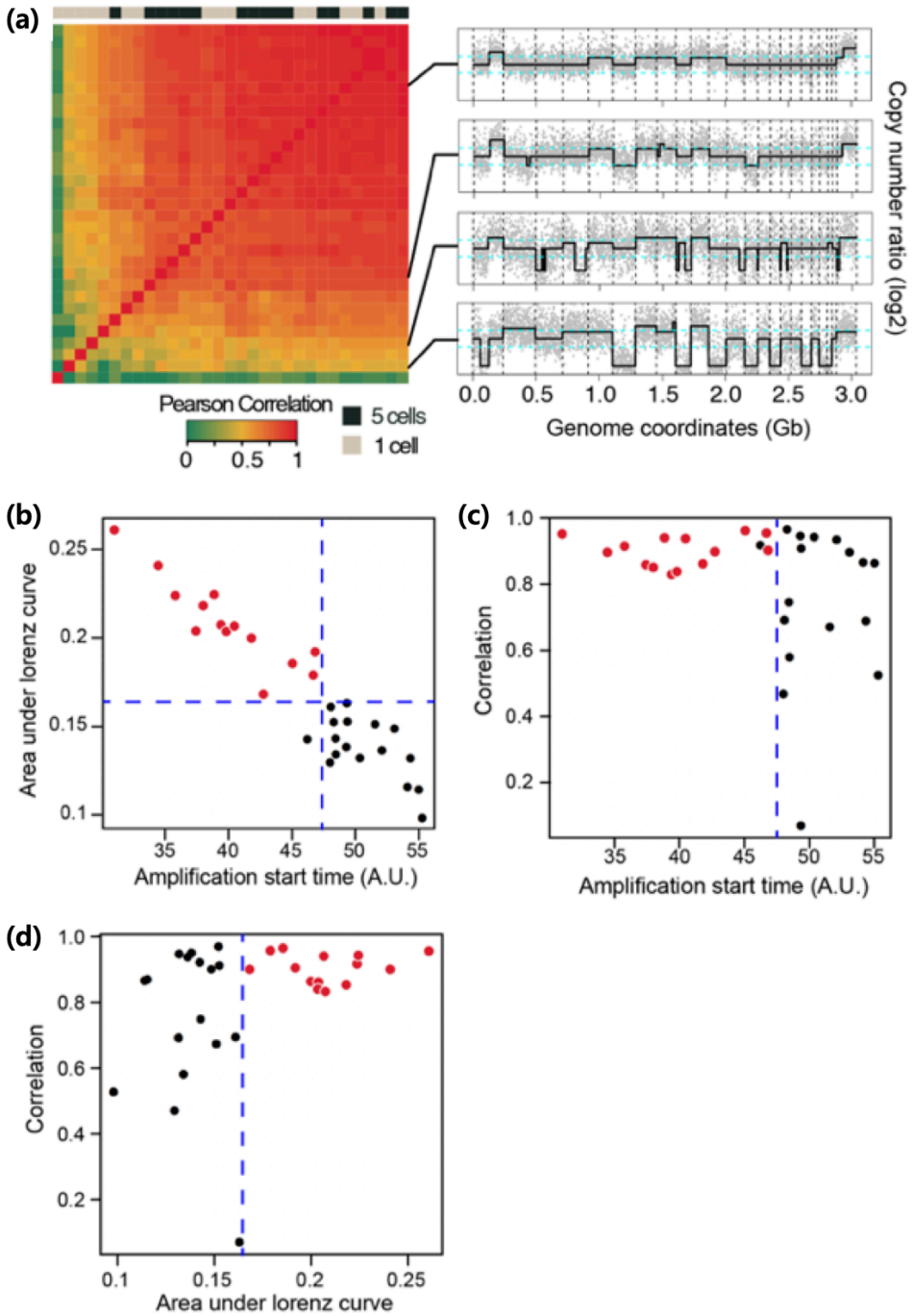

Figure 2.10 High genomic coverage and a high correlation with the true copy number value of the whole-genome amplified samples were related to low C_T values for real-time monitoring of the amplification reactions [25]

the 16 primer panels to enrich the high-quality amplified products before sequencing.

In addition, it was tested whether the irradiating IR laser pulse and vaporizing discharging layer generate DNA fragmentations or not. For this purpose, single HL60 cells were exposed to 0 to 50 shots of laser pulses. However, we could not detect any signs of fragmentation in the sequencing data, even from cells that had repeatedly been irradiated by the IR laser pulse (Figure 2.11).

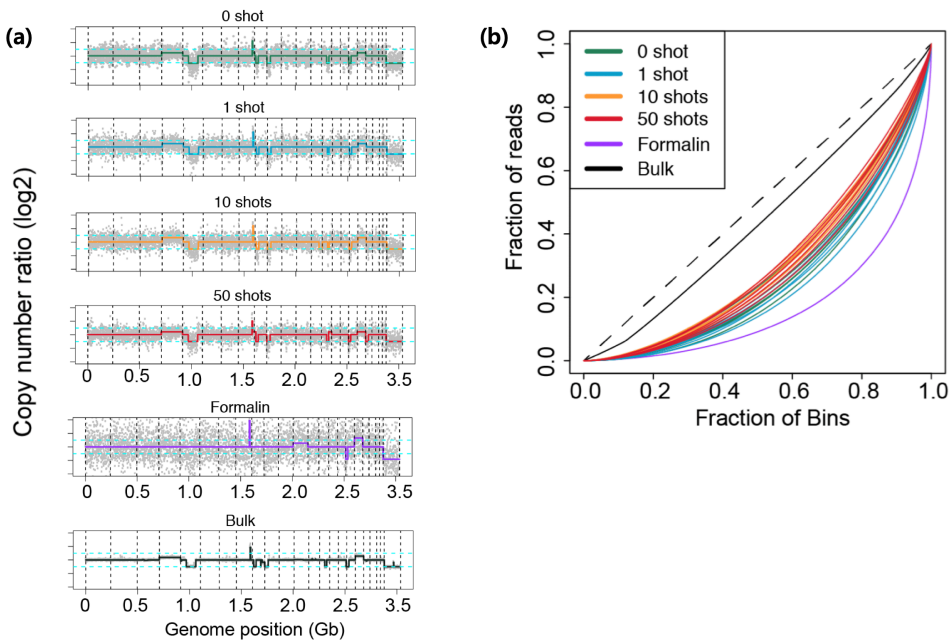

Figure 2.11 Single HL60 cells were exposed to 0 to 50 shots of laser pulses to test whether the irradiating IR laser pulse and vaporizing discharging layer create damages on cells or not [25]

To confirm the sequencing performance, HL60 cells were isolated using the PHLI-seq method, sequenced their genomes, and tested the coverage breadth, allele dropout (ADO), and false positive rate (FPR). To characterize and validate the quality of PHLI-seq according to whole genome amplification methods and the number of input cells, single or twenty cells' whole genomes of a cell line were amplified by multiple displacement amplification (MDA) or multiple annealing and looping based amplification cycles (MALBAC). First, coverage breadth and uniformity of each group were evaluated. It was found that coverage breadth in twenty-cell group (88.18%± 0.70%) was almost equivalent to that of the bulk population (90.92%) (Figure 2.12). In contrast, single-cell group showed limited coverage (77.8%±3.30%) compared to twenty-cell group (P=0.0079, Wilcoxon rank sum test).

On the other hand, single-cell group showed 18% to 40% of allele dropout (ADO) rate. Also, ADO rate was lower for MDA compared to MALBAC. In contrast, much lower ADO rates were observed in twenty-cell group (P = 0.0285, Wilcoxon rank sum test), showing 0.956% on average, which is 32.3-fold reduced dropout rate compared to the single-cell group (Figure 2.13a). False positive rate (FPR) in single-cell and twenty-cell groups, which were amplified by MDA were 2.13×10^{-5} and 5.65×10^{-6} respectively. However, MALBAC showed much higher FPR in both groups (8.4×10^{-4} and 5.1×10^{-5} from single-cell and 20-cells group, respectively). This could be attributed to BstI polymerase used in MALBAC in that the polymerase does not have proof-reading activity.

Figure 2.13b represents that variant allele frequency distributions. In that figure, you can notice that only reactions from twenty cells amplified by MDA had similar distribution with that of gDNA from population of cells. Also, the distribution from population of cells was normally distributed, but amplified products presented bimodal or skewed distribution. This result could be ascribed to amplification bias which makes detected allele frequency of a locus deviate from the original value, and polymerase error which creates a new population in a low range of allele frequency. Taken together, it was concluded that amplification of approximately 20 cells by MDA is appropriate to reduce the effects of amplification bias and polymerase error.

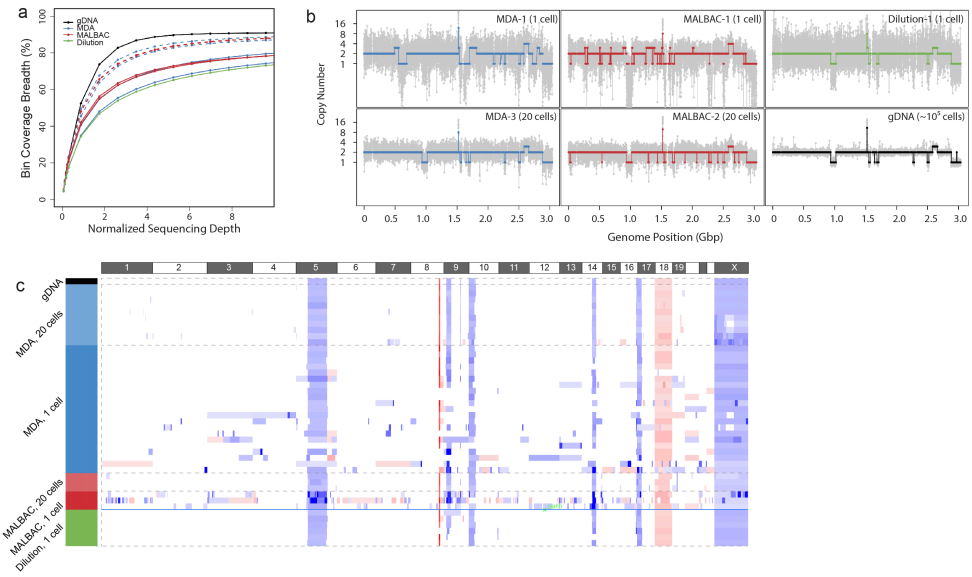

Figure 2.12 Validating sequencing quality of PHLI-seq. HL60 cells were isolated by PHLI-seq and whole genome amplified by MDA or MALBAC. (a) Bin coverage breadth according to sequencing depth and preparation methods. Dashed lines and solid lines represent 20-cell and single-cell isolation, respectively. Sample 'Dilution' indicates sequencing result of single-cells isolated by dilution method and amplified by MDA. (b-c) Copy number plot of the analyzed samples [25].

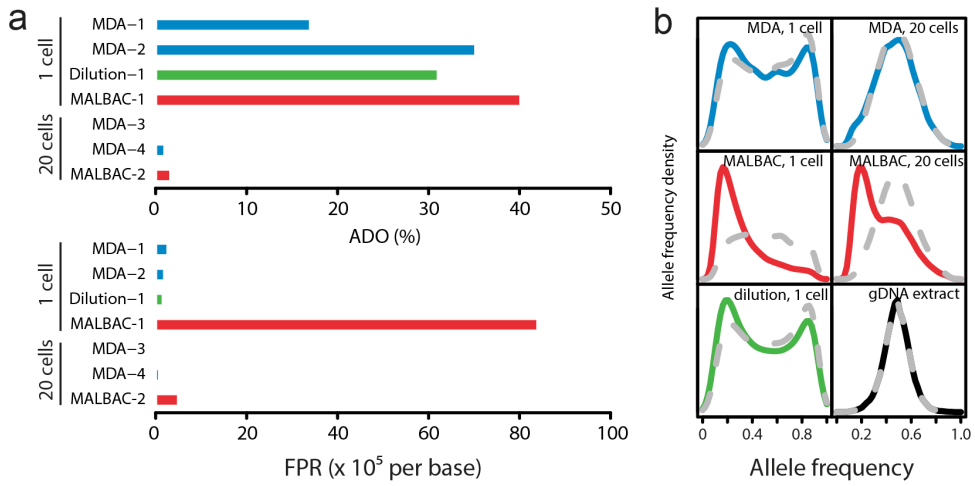

Figure 2.13 Validating sequencing quality of PHLI-seq. HL60 cells were isolated by PHLI-seq and whole genome amplified by MDA or MALBAC. Amplified DNA was analyzed whole exome sequencing (WES). Then, allele dropout (ADO) and false positive rate (FPR) were calculated. (b) Allele frequency distribution according to preparation methods [25]

2.5 Performance Comparison with Commercial Laser Microdissection

In this section, commercially available LMD and laser pressure catapulting (LPC) techniques will be compared to PHLI-seq (Figure 2.14). LMD technique uses UV pulsed laser to dissect a region of a sample and the polymer membrane which supports the region. LPC technique also uses UV laser but catapults a cell by pressure generated by laser-produced plasma. In using LMD or LPC, UV laser-induced damage, electrostatic adhesion of a dissected sample, and slow process could be potential problems. Therefore, we evaluated whole-genome sequencing qualities of single HL60 cells through LMD, LPC, and PHLI-seq techniques. We used Leica LMD6500 (Leica Microsystems, Wetzlar, Germany) to perform LMD and LPC for cell isolation. We used nuclease and human nucleic acid free PET-membrane FrameSlide (Leica Microsystems) to prepare samples. Laser power was adjusted down to a minimum level (40~45 in the operating software for LMD6500) to minimized UV damage to cells. For LPC, the aperture in the instrument was opened to maximum and the laser was out-focused. Post-processing for LMD and LPC is same with PHLI-seq.

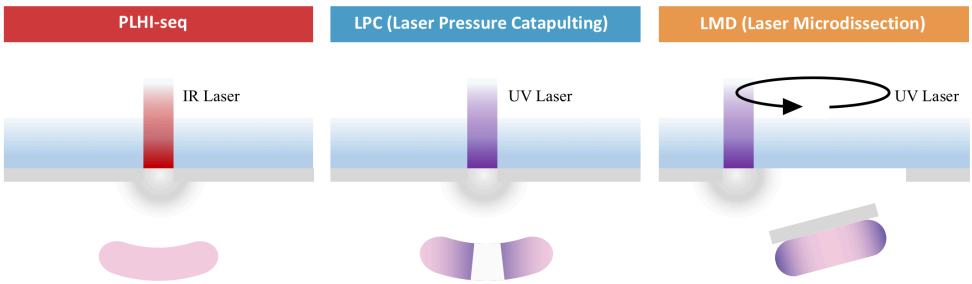

Figure 2.14 Methodological comparison of commercially available laser microdissection techniques and PHLI-seq

For comparison, a cell line was used rather than a tissue section, because cells in a tissue section could have partial genome by tissue sectioning, which is not suitable for a well-controlled experiment. After preparing cells on substrates, cells were isolated through LMD, LPC, and PHLI-seq techniques. Isolated cells were lysed and their whole-genome was amplified by MDA. Since the MDA mechanism hinders damaged or fragmented genome to be amplified thoroughly, we monitored the MDA procedure with a real-time PCR machine so that we can measure amplification start time as done with the cell line validation experiments. As represented in the Figure 2.15a, it was found that the cells isolated by PHLI-seq showed significantly earlier amplification start time with smaller deviation compared to LPC and LMD (Wilcoxon rank sum test, $p < 10^{-3}$). Therefore, it was concluded that PHLI-seq can produce higher quality of sequencing data than LPC and LMD.

In addition, throughput of cell isolation was compared for the three techniques. The time from targeting the cells at the computer interface to when the cells were completely isolated was measured (Figure 2.15b). The result shows that LMD is

much slower than other two techniques (Wilcoxon rank sum test, $p < 10^{-3}$). Since PHLI-seq and LPC can isolate cells through single laser pulse, the two techniques are relatively fast procedure. While LMD must dissect the contour of the region of interest, it takes longer time than PHLI-seq and LPC. For LMD, there was also a problem that samples could not be easily detached from slides even after complete dissection.

The amplified DNA samples were analyzed by low-depth WGS. Using the sequencing data, 'alignment ratio' and 'correlation' were analyzed to get an insight on overall sequencing quality. 'alignment ratio' is the portion of reads aligned to the human genome and 'correlation' is chromosomal copy number correlation between the sequenced sample and a population of HL60 cells (Figure 2.16). From the plot, several samples that were isolated by LPC and LMD had low alignment ratio (< 0.2). Furthermore, the single-cells, which were isolated by PHLI-seq, showed higher correlations to the true HL60 genome compared to those isolated by LPC and LMD. This means that dissected or catapulted cells through LPC or LMD were not collected into PCR tubes, or isolated nucleic acids were severely damaged to be amplified.

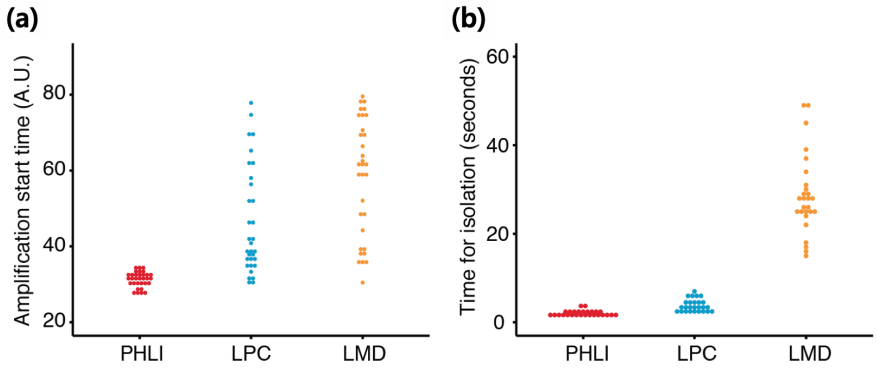

Figure 2.15 PHLI-seq, LPC, and LMD were compared from the perspective of 'Amplification start time' and 'Time for isolation' [25]

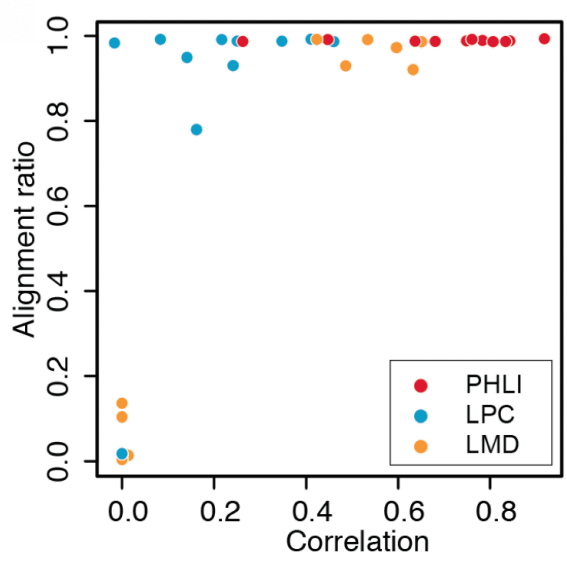

Figure 2.16 PHLI-seq, LPC, and LMD were compared in the 'Alignment ratio' vs. 'Correlation' plane. Here, the higher alignment ratio is, the more NGS reads are aligned to reference genome sequences. Correlation means chromosomal copy number correlation between the sequenced samples and genomic DNA of a population of HL60 cells [25]

As aforementioned, the uniformity of the amplification also indicates the quality of isolated samples. In other words, the uniform amplification of a sample can indicate that the cell has been detached without damage or fragmentation, and the Lorenz curve could provide an insight on the DNA quality after isolating cells by PHLI-seq, LPC, and LMD (Figure 2.17a,b). The results indicate that DNA quality is highest when cells were isolated through PHLI-seq. Then, two samples with highest qualities were selected for each isolation method, and their genome-wide copy numbers were analyzed through NGS (Figure 2.17c). The CNA plots from PHLI-seq technique showed uniformly amplified genome with correct copy number except for single-cell variability. In contrast, LPC and LMD presented large deletions or severe amplification bias, which clearly showed technical superiority of PHLI-seq.

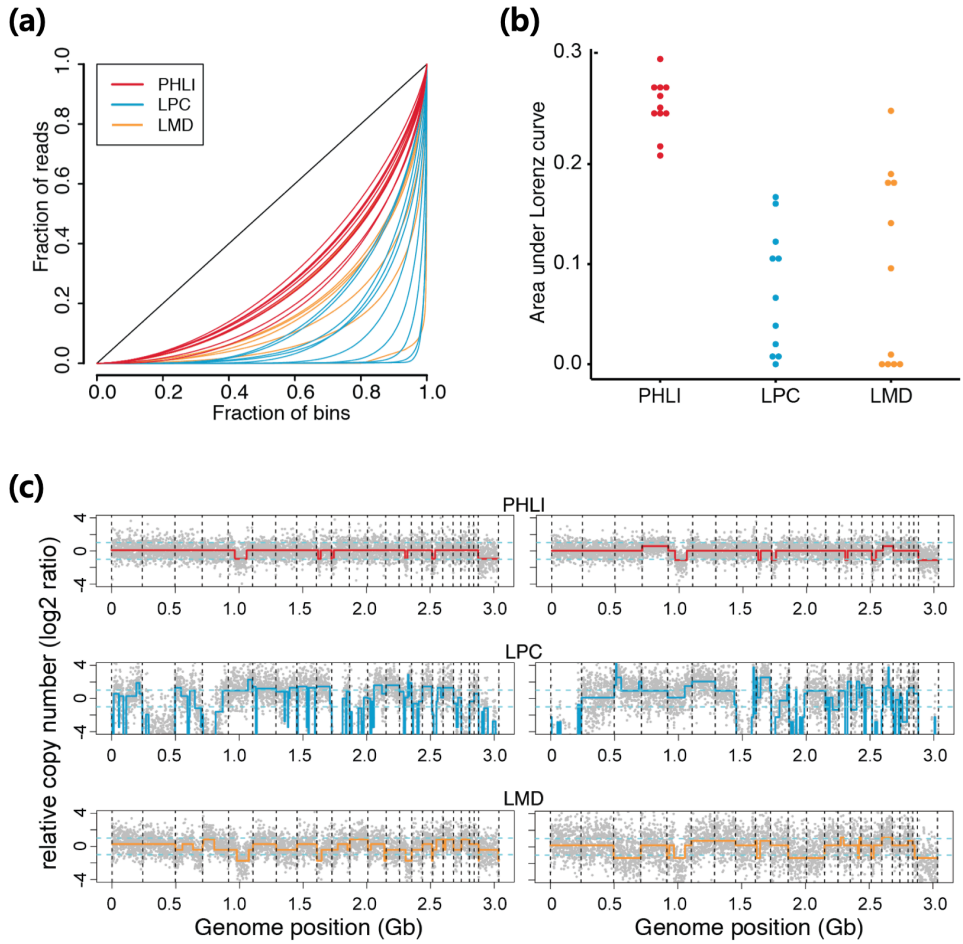

Figure 2.17 Lorenz curves and copy number plots were drawn for the samples which were isolated by the three methods. (a,b) Lorenz curve quantitatively shows overall uniformity of whole genome amplification. (c) Copy number plots show qualitative uniformity and correctness of chromosomal copy numbers [25]

Chapter 3

Spatially Resolved Sequencing of Breast Cancer

In this chapter, two breast cancer cases analyzed by PHLI-seq are introduced. The first case is a hormone receptor (HR)-positive/ Human epidermal growth factor receptor 2 (HER2)-positive invasive ductal carcinoma (IDC) to demonstrate spatially resolved sequencing of breast cancer. Based on an inspection by pathologists, 53 cell clusters were selected in the tissue section for analysis by PHLI-seq. In the second case, consecutive sections of a triple-negative (estrogen/progesterone receptor and HER2-negative) breast cancer sample underwent PHLI-seq analysis to discover how heterogeneous tumor subclones exist in the three-dimensional space of the tissue and to demonstrate how PHLI-seq can be an empowering tool to bridge genomics to histopathology.

3.1 PHLI-seq Analysis of a HER2 Positive Breast Tumor

3.1.1 Experimental Procedure

A fresh-frozen breast cancer tissue was obtained from the Department of Surgery, Seoul National University Hospital. The sample was analyzed under the approved Institutional Review Board (IRB) protocol (IRB No. 1207-119-420). The tissue used in this experiment was obtained from a 57-year-old woman who underwent total mastectomy with axillary lymph node dissection in April 2014. The breast cancer was a Stage IIIC (pT2N3M0, AJCC 7th TNM Staging) IDC which was positive for estrogen receptor (95%), progesterone receptor (2%), and HER-2 (+++/3) on IHC, as evaluated according to the American Society of Clinical Oncology and College of American Pathologists (ASCO/CAP) guidelines**.** The frozen breast cancer tissue was stored at −80 °C until it was sliced into 10-μm-thick sections. The tissue section was thaw-mounted onto ITO-coated glass slides. After mounting, the tissue sections were dried for 15 minutes at room temperature and subjected to a modified hematoxylin and eosin (H&E) staining technique. The protocol was as follows:

1) rinse in tap water for 5 minutes

2) stain in Harris haematoxylin solution (Merck, Darmstadt, Germany) for 3 minutes

3) rinse in tap water (quick dip)

4) rinse in 1% HCl solution/EtOH (quick dip)

5) rinse in tap water for 5 minutes

6) counterstain with eosin Y (BBC Biochemical, Mount Vernon, WA) for 3 seconds

7) rinse in water (10 dips)

8) dehydrate in 70% EtOH (10 dips)

9) dehydrate in 90% (10 dips)

10) dehydrate in 100% EtOH (10 dips).

The sections were allowed to dry at room temperature. Prepared glass slides were stored at −20 °C until analysis. Then, the prepared sections were scanned using an automated microscope to generate a high-resolution whole-section image. A total 53 cell clusters were selected based on an inspection by pathologists, isolated sequenced by PHLI-seq technique (Figure 3.1), and their whole-genomes were amplified by MDA.

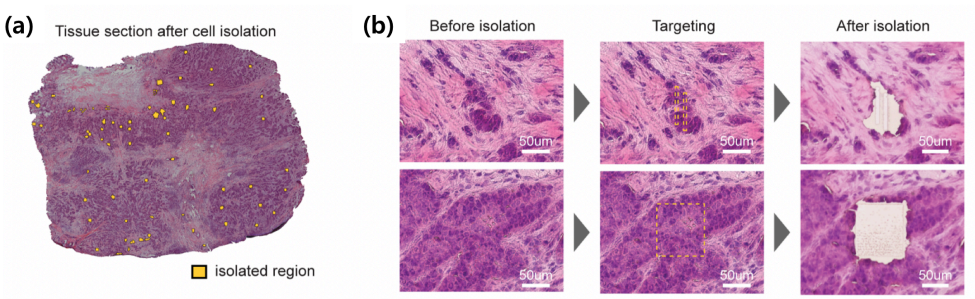

Figure 3.1 (a) The breast tumor tissue section was imaged after cell isolation (b) Regions of before and after tumor cell isolations by the PHLI-seq instrument [25].

The whole-genome amplified products were fragmented using an EpiSonic Multi-Functional Bioprocessor 1100 (Epigentek) to generate a 150~250-bp fragment distribution. The fragmented products underwent sequencing library preparation for Illumina platform. In brief, end repair, 3'dA-tailing, adaptor ligation, and PCR

amplification according to the manufacturers' instructions were performed. For sequencing library preparation, Celemics NGS Library Preparation Kit (LI1096, Celemics, Seoul, Korea) for the whole-genome sequencing library preparation, SureSelectXT (Agilent, CA, US) for whole-exome sequencing, and the Celemics Customized Target Enrichment Kit (SICT96, Celemics, Seoul, Korea) for targeted sequencing were purchased and used as described in manufacturer's manuals. DNA purification was performed by ^{TOPQ}XSEP MagBead (XB6050, Celemics, Seoul, Korea), and DNA libraries were amplified using the KAPA Library Amplification Kit (KAPA Biosystems, KK2602). Final library products were quantified by TapeStation 2200 (Agilent, CA, US). For sequencing run, HiSeq 2500 50SE (Illumina) was used to generate 0.16 Gb/sample for whole-genome sequencing. Also, HiSeq 2500 150PE (Illumina) was used to generate 5 G/sample and 0.88G/sample for whole-exome and targeted sequencing, respectively.

The NGS sequence reads were mapped to the GRCh37 human reference genome using BWA-MEM [36] (version 0.7.8) with default parameters. The resulting SAM files were sorted by chromosome coordinates. Then, PCR duplications were marked through using Picard (version 1.115) (http://picard.sourceforge.net/). Reads with a mapping quality score less than 30 or that have a supplementary alignment were removed from the BAM file.

3.1.2 Somatic Copy Number Alteration (CNA) Analysis

For CNA analysis, low-depth whole-genome sequencing data were utilized. The CNAs of the samples were estimated using variable-size binning method [37]. In brief, this method first divides the region of whole genome into 10,000 variable-sized bins, in which each bin has an equal number of expected uniquely mapped

reads (median genomic length of bin = 276 kbp). For this experiment, 7,000 bins (median = 396 kbp) were created. Then, each NGS sequence read was assigned to each bin followed by Lowess GC normalization to obtain the read depth of each bin. Chromosomal copy numbers of a sample wree estimated by normalizing the read depth of each bin by the median read depth of the reference DNA.

After the CNAs detection followed by applying MergeLevels [38], the data underwent multi-sample segmentation with gamma=20. Given the multi-sample segmentation, the event vector was constructed for each cell by providing a value of 0 for the segments with a median copy number of reference DNA, 1 for segments with a gain, and -1 for segments with a loss. Finally, a correlation matrix was constructed, and hierarchical clustering was performed. Then, a multiscale bootstrapping resampling method was applied to calculate the unbiased p-value [39]. This step was for evaluating the accuracy of the clustering. The approximately unbiased (AU) p-value indicated the strength of the cluster supported by bootstrapping. After creating bootstrap data, the data were clustered repeatedly. The AU p-value of a cluster was related to the frequency that appeared in the bootstrap replicates, and the method was implemented by the R package Pvclust [40].

Genome-wide CNA analysis revealed three subclonal populations existed in the tumor sample (Figure 3.2, approximate unbiased p-value > 0.99, multiscale bootstrap resampling with 10,000 iterations). Generally, the three subclonal populations had similar CNA profile with unique variants in several regions in chromosomes. The shared alterations include 1q gain, 8q gain, 8p loss and HER2 amplifications, all of which had been previously reported as frequent CNAs in

human breast cancer and other types of cancer [26, 27]. One interesting observation would be the fact that the CNAs status was clearly divided into three distinct populations without intermediate subclones between the three types. In this sample, the causative fitness gains underlying breast carcinogenesis appeared to take place in the *MYC* (8q24), *AKT3* (1q44), or *ERBB2* (HER2) (17q12) regions, all of which were gained or amplified in 12.4%~21.3% of cases in The Cancer Genome Atlas (TCGA) project. The up-regulation of these genes are known to be related to the cause of breast cancer [41]. In contrast, subclones 1, 2 and 3 harbor unique gains or focal amplifications at 6q22.32-q23.2, 2q14.1-q14.2 and 8p21.3-p22, respectively. Among them, the amplification of *CUG2* (Cancer-Upregulated Gene 2, also called *CENPW*, 6q22.32) which is known to be associated with tumor progression [43] might contribute for tumor cells included in subclone 1 to proliferate aggressively.

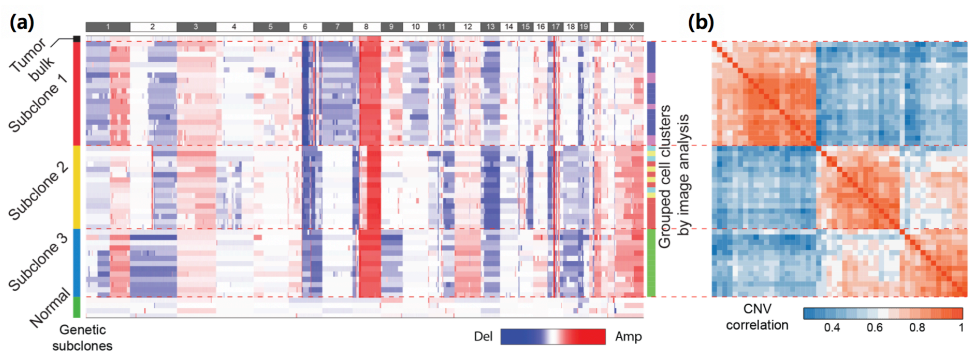

Figure 3.2 (a) The heatmap of CNAs through low-depth whole-genome sequencing. Rows were reordered to cluster samples by the correlation of their CNAs. Three distinct genetic subclones were identified by this analysis. The three subclones had both shared and exclusive CNA events. (b) The correlation matrix of the copy number data [25]

After finding the three subclones we isolated additional cell clusters (n=27) at the boundaries between subclones, since intermediate subclones might be excluded from the sampling process. The isolated samples were analyzed by similar manner. Clustering analysis was performed based on the inferred copy number data for both previously isolated (n=53) and additionally isolated samples (n=27). The results showed that the 80 cell clusters from the tumor sections were classified into one of the three previously defined cancer subclones (Figure 3.3). This result reinforces that there was no intermediate subclone and supports the theory of punctuated copy number change followed by a period of stasis, as demonstrated in previous studies [16, 18, 28].

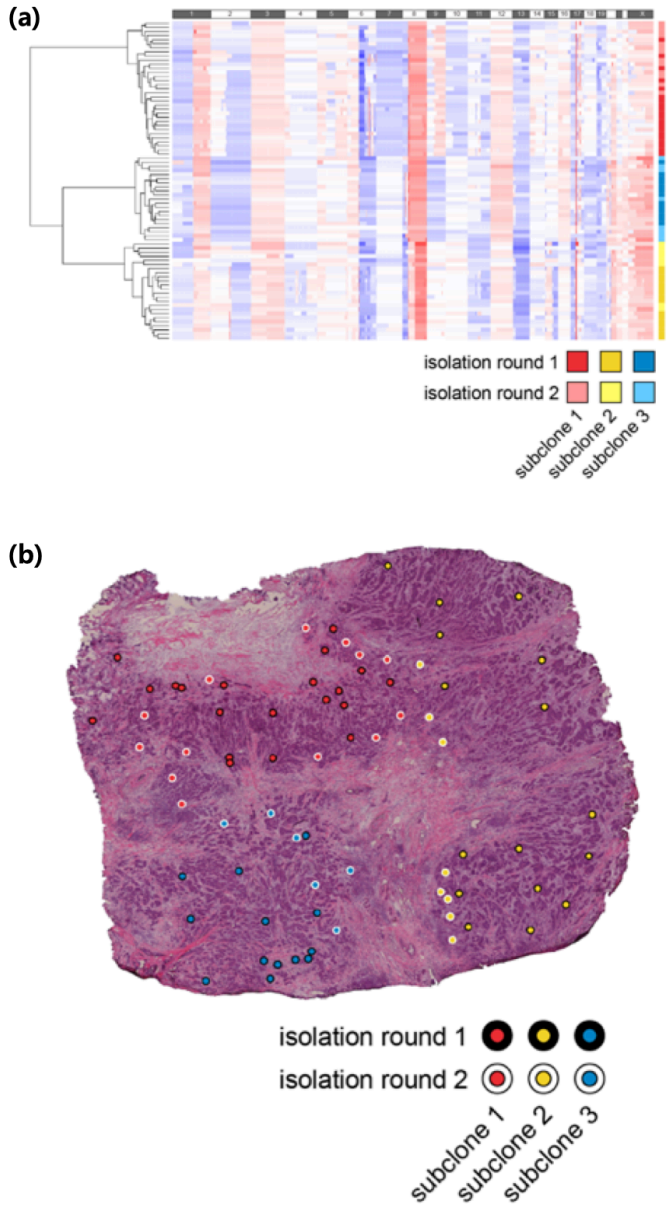

Figure 3.3 Additional cell clusters (n=27) were isolated at the boundaries between subclones. The isolated samples were analyzed by low-depth whole-genome sequencing. Then, clustering analysis showed that the 80 cell clusters from the HER2-positive tissue sections were classified into one of the three previously defined cancer subclones [25].

3.1.3 Somatic Single Nucleotide Variant (SNV) Analysis

Before SNVs detection, GATK (v3.5-0) IndelRealigner and BaseRecalibrator were used to locally realign reads around the Indel and recalibrate the base quality score of BAM files [46]. Then, three different variant callers (GATK UnifedGenotyper, Varscan, and MuTect) were used to detect somatic SNVs to avoid false-positive variant detection [47]. GATK UnifiedGenotyper was used with default parameters.

Next, GATK VariantRecalibrator to obtain filtered variants [46]. Tumor bulk sample and PHLI-seq sorted sample data were processed together to produce a single vcf file. The training data used for variant recalibration included dbSNP build 137, hapmap 3.3, Omni 2.5, and 1000G phase1, and QD, MQ, FS, ReadPosRankSum, and MQRankSum annotations were used for the training. Variants detected in the paired blood sample were considered as germline variants, which were removed from variants to produce the final list of GATK called variants. Varscan2 [48] (ver 2.3.7) and Mutect [49] (ver 1.1.4) were used with default parameters to produce the lists of Varscan and MuTect called variants, respectively.

Among the variant loci detected in the samples by the three variant callers, ‘intra-sample double called sites’ were defined as loci where at least two callers detected variants. By considering only these variants for subsequent analysis, we could reduce false-positive variants [47]. Among the intra-sample double called sites, variants detected in only one sample were excluded in analysis to remove WGA errors (generated by polymerase during ampfliciation), and the genomic loci with the resultant variants were considered confident sites. Finally, a variant in the confident sites was considered to be true variant in the case of one of the three

variant callers detected the variant at the locus and the allele count of the variant was significantly larger than that of the other non-reference bases (Fisher's exact test, $p < 10e\text{-}4$).

To investigate somatic SNVs more sensitively, we carried out targeted sequencing of 121 genes associated with breast cancer (Table 2). 121 genes were chosen from the SNUH BCC (Seoul National University Hospital Breast Care Center) Panel. The targeted sequencing revealed unique mutational profiles in each subclone (Figure 3.4). In the targeted sequencing analysis of 53 cell cluster samples, it was found that mutations in *PIK3CA*, *EPHA3*, *KIT*, *ERBB4* and *KMT2C* occurred in subclone 1, mutations in *KMT2C*, *ATR* and *KDM5B* in subclone 2, and mutations in *PIK3CA, TOP2A*, *NF1*, *ESR1*, *JAK1* and *MST1R* in subclone 3. Especially, the mutation at PIK3CA p.M1043I, which was detected both in subclone 1 and 3, has previously been known to be an oncogenic mutation that causes an increase in PI3K lipid kinase activity, constitutive AKT activation and the transformation of NIH3T3 cells and chick embryo fibroblasts [4, 5].

For further analysis, whole-exome sequencing was carried out for four samples selected from each subclone (Figure 3.5). 75 mutations were shared in the three subclones and 99, 75, and 382 mutations occurred exclusively in subclones 1, 2, and 3, respectively. Interestingly, most of the mutations detected by PHLI-seq in the subclones 3 were not detected in the sequencing data of the tumor bulk. This means that PHLI-seq can provide rich information about subclonality and variants with a low-level allele fraction in heterogeneous tumors.

Figure 3.4 Somatic SNV analysis through targeted sequencing of 121 genes associated with breast cancer (Table 2). The 121 genes were chosen from the SNUH BCC (Seoul National University Hospital Breast Care Center Panel [25].

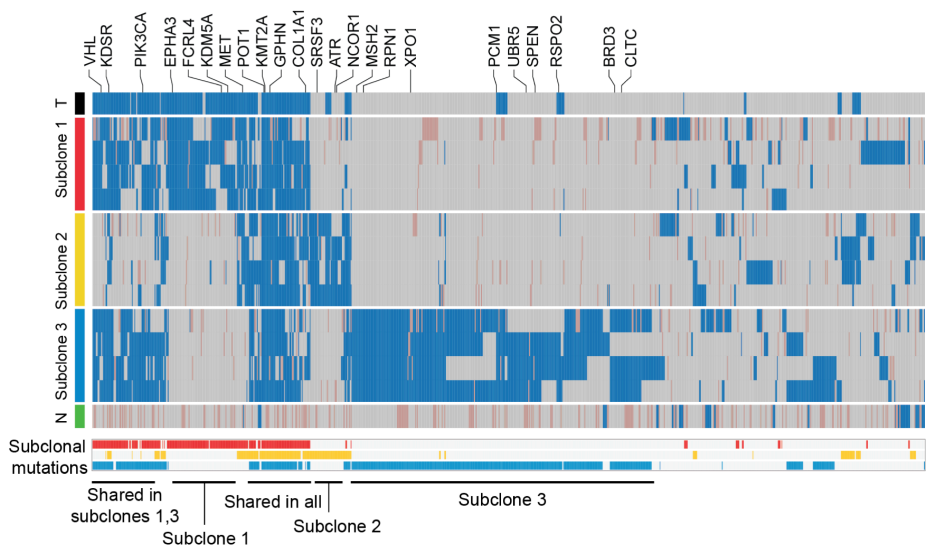

Figure 3.5 Somatic SNV analysis through whole exome sequencing [25].

Table 2 The list of 121 genes associated with breast cancer, which were used for targeted sequencing experiment.

MTOR	SETD2	IGF2R	NOTCH1	MDM2	NCOR1
EPHA2	MST1R	EGFR	GATA3	NAV3	NF1
ARID1A	EPHA3	CDK6	RET	TBX3	CDK12
PIK3R3	POLQ	PIK3CG	PTEN	FLT3	ERBB2
JAK1	ATR	MET	FGFR2	BRCA2	TOP2A
NOTCH2	PIK3CA	BRAF	MEN1	RB1	BRCA1
MCL1	FGFR3	EZH2	MALAT1	IRS2	RPS6KB1
DDR2	PDGFRA	KMT2C	CCND1	FOXA1	STK11
ABL2	KIT	ZNF703	FGF4	MAP4K5	INSR
KDM5B	INPP4B	EIF4EBP1	FGF3	AKT1	TYK2
IKBKE	FBXW7	FGFR1	C11orf30	LTK	NOTCH3
PARP1	MAP3K1	PRKDC	PAK1	IDH2	JAK3
AKT3	PIK3R1	MYC	MRE11A	IGF1R	CCNE1
ALK	APC	PTK2	ATM	TSC2	AKT2
SF3B1	PDGFRB	JAK2	FOXM1	CBFB	SRC
IDH1	FGFR4	CDKN2A	CDKN1B	CTCF	AURKA
ERBB4	FLT4	CDKN2B	KRAS	CDH1	GNAS
FANCD2	DDR1	SYK	KMT2D	TP53	PTK6
VHL	ROS1	TLR4	ERBB3	AURKB	RUNX1
CTNNB1	ESR1	ABL1	CDK4	MAP2K4	EP300
AR					

Although the stringent variant calling pipeline was used to detected somatic SNVs, the SNVs should be verified through single-molecule deep sequencing [52], [53], since initial nucleic acids were amplified by whole genome amplification. For this, a portion of mutations that were detected using PHLI-seq was targeted, and sequencing libraries were generated for the targeted sites by tagging a unique molecular barcode to each DNA molecule to precisely discriminate the next-generation sequencing results from errors (Figure 3.6).

The difference from the original paper is that a targeted library was generated using PCR with minimal cycles before tagging the molecular barcodes. This would lose the ability to call duplex consensus sequences (DCSs), but it is still possible to call single-strand consensus sequences (SSCSs). 3-plex PCR using 50 ng of genomic DNA was carried out with a 2X KAPA2G Fast Multiplex kit (KAPA Biosystems, KK5802). The amplified products underwent gel electrophoresis, gel purification, and quantification by a Qubit high-sensitivity dsDNA quantification kit (Invitrogen), and normalization for pooling. The pooled library was end-repaired, dA-tailed (NEB, NEBNext® End Repair Module, E6050S), and ligated with duplex tags. Then, DNA molecules with a length of 200 to 1000 bp were purified to remove dimers. Finally, the molar concentration of the tag-ligated libraries was quantified. The constructed library was sequenced using MiSeq 250 PE (Illumina) to generate a 18,182X (median) single-molecule sequencing depth. After generating the SSCSs, the targeted loci (where SNV had been detected) and background loci (where no SNV had been detected) were split.

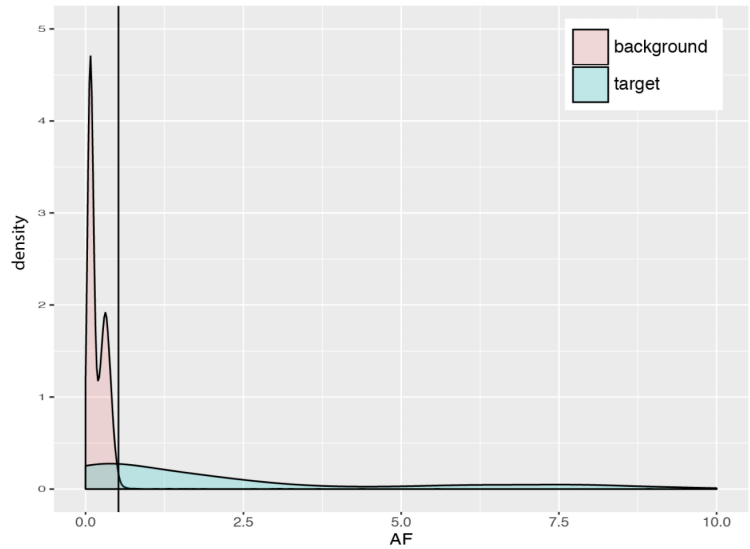

Figure 3.6 Single molecule deep sequencing result. From the sequencing result, single strand consensus sequence (SSCS) was identified, and SNVs detected by PHLI-seq were validated. SNVs detected in PHLI-seq were grouped to 'target', and those were not detected to 'background'. Two groups had different allele frequency distribution, and a threshold for true SNV was set to guarantee Benjamini–Hochberg false discovery rate < 0.05 (horizontal line) [25].

The validation limit was set to 0.55% to limit the false discovery rate to < 0.05. The results validated 92.3% (12/13) of the subclonal mutations detected by targeted sequencing were true somatic variants (Figure 3.7). Moreover, 72.2% (78/108) of the mutations observed in the whole-exome sequencing were validated. The validation rate was 90.9% (50/55) when mutations that were observed in more than three samples were considered.

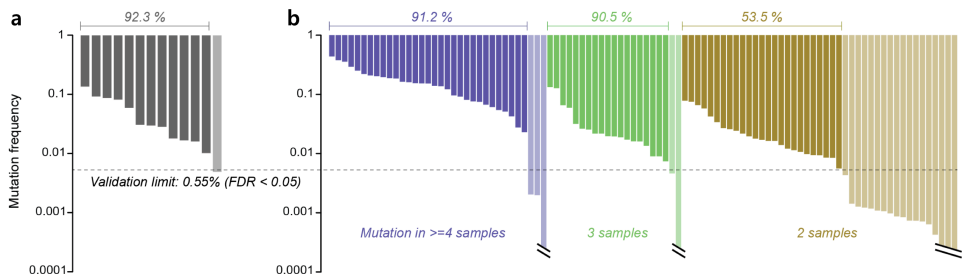

Figure 3.7 Mutation validation using single-molecule deep sequencing. A portion of mutations that were detected using PHLI-seq were selected. Then, sequencing libraries for the targeted sites were constructed by tagging a unique molecular barcode to each DNA molecule to precisely discriminate the next-generation sequencing results from errors. The validation limit was set to 0.55% to limit the false discovery rate to < 0.05 [25].

3.1.4 Inferring Tumor Evolution

We have mapped the detailed information for CNAs, driver mutations and passenger mutations to the topological information and the spatial positions of the tumor tissue. It was found that the three subclones were spatially separated in the tumor mass. As shown in Figure 3.8a, whereas the heterogeneity of the tumor tissue is clear from the detection of the three different subclones, the micro area occupied by each subclone exhibits no mingling with cells from other subclones. This finding implies that the three subclones are independent and have established tumorigenic advantages, suggesting that a combination of different drugs to inhibit different subclones in a patient should be a future therapeutic strategy for personalized cancer medicine.

Based on the CNA and SNV analyses, we were able to infer the evolutionary history of the subclones in the tumor (Figure 3.8b). A precursor clone could be generated by obtaining an advantage for tumor proliferation by accumulating CNAs and mutations that are shared by the three subclones. However, there could be the possibility that the genesis of CNAs and mutations shared by the three subclones might have been completed over several generations of previous initial subclones originating from a *bona fide* ancestor clone, giving rise to the birth of a plausible precursor clone that harbored the variants shared by the three subclones. Then a cell in the original precursor clone might have obtained other variants, which became a subsequent precursor clone for the generation of subclones 1 and 3. Other cells in the original precursor clone could accumulate different variants, which could lead to subclone 2, which branched off evolutionarily from subclones 1 and 3. This result can be inferred from the

observation that subclones 1 and 3 have many common mutations, including PIK3CA p.M1043I, whereas subclone 2 has exclusive mutations and mutations that are common to all three subclones.

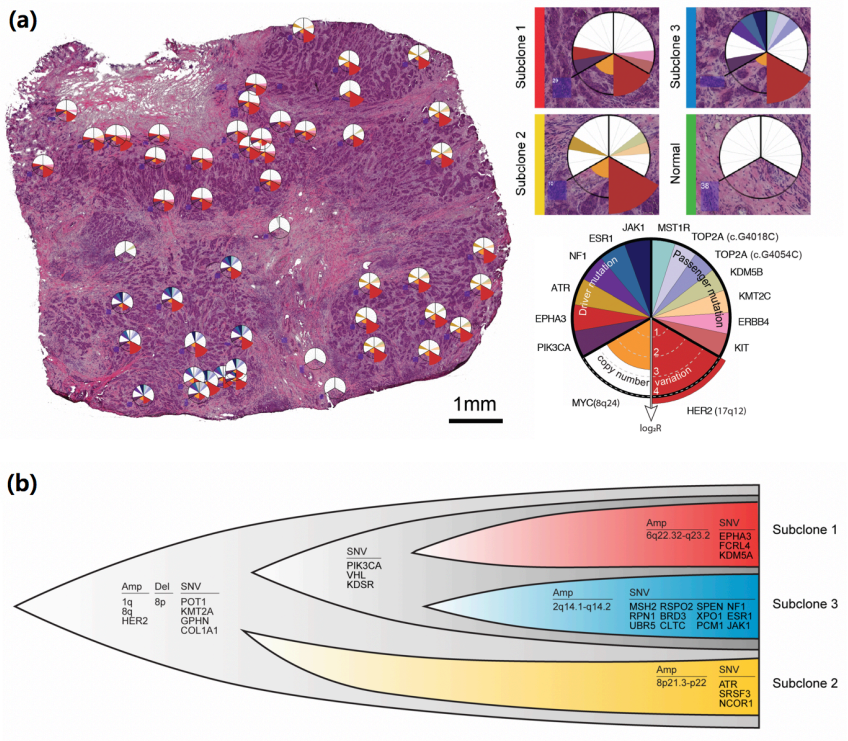

Figure 3.8 Spatially resolved subclonal structure and their evolutionary history was inferred based on sequencing and histologic information. Alterations shared by the three subclones may have led to the early stage of tumorigenesis. Subclone 2 and the ancestor of subclones 1 and 3 may have divided and accumulated unshared alterations [25].

In addition, a copy number analysis of the *HER2* gene showed that subclone 2 has a much higher *HER2* amplification level than subclones 1 and 3, which further supports this scenario (Figure 3.9). After being split into the two lineages, the precursor subclones would have accumulated copy number changes or mutations independently, and one of the two could be further split to ultimately give rise to subclones 1 and 3. Subclone 3 might have accumulated more mutations, possibly due to the exclusive mutation in the MSH2 gene encoding a component of the post-replicative DNA mismatch repair system.

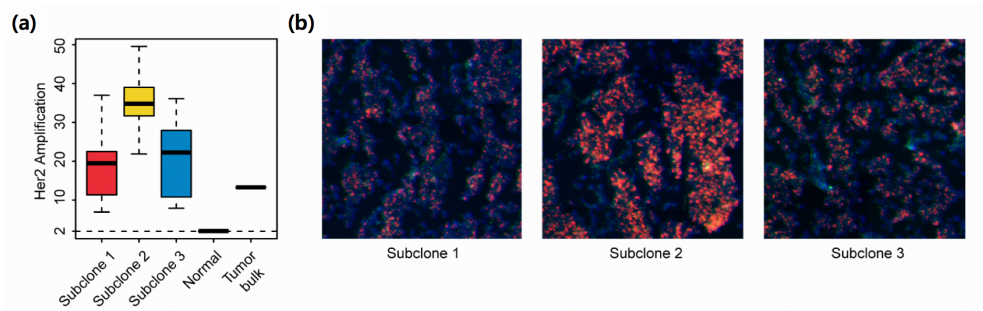

Figure 3.9 (a) HER2 gene copy numbers in each subclone were identified by whole genome sequencing analysis and (b) their fluorescent in situ hybridization staining [25]

3.2 PHLI-seq Analysis of a Triple-negative Breast Tumor

As an additional case, consecutive sections of a triple-negative (estrogen/progesterone receptor and HER2-negative) breast cancer sample underwent PHLI-seq analysis to discover how heterogeneous tumor subclones exist in the three-dimensional space of the tissue and to demonstrate how PHLI-seq can be an empowering tool to bridge genomics to histopathology (Figure 3.10). The tissue used for 3D mapping was obtained from a 56-year-old woman who underwent total mastectomy with axillary lymph node dissection. The breast cancer was a Stage IIA (pT2N0M0, AJCC 7th TNM Staging) IDC which was negative for estrogen receptor (1%), progesterone receptor (negative) and HER-2 (-/3) on IHC. Size of the tumor was approximately 7 x 6 x 5 mm, and seven tissue sections, each 700 um apart, were used to prepare H&E sections for PHLI-seq analysis. A total of 177 cell clusters from the seven H&E sections were isolated and sequenced by PHLI-seq as described. Before the isolation, cancer cells with various phenotypes were identified by histopathological evaluation from H&E and IHC sections (Figure 3.11). Based on the phenotypic information, the 177 cell clusters were selected to measure genetic heterogeneity in tumor cells with different phenotypes. After isolating the targeted cell clusters, low-depth whole-genome sequencing was performed to obtain an image of the heterogeneity of CNAs (Figure 3.10). Through the sequencing result, three genetic subclones were observed and they were denoted *in situ* clone 1, *in situ* clone 2, and invasive clone. *In situ* clone 1 and *in situ* clone 2 had the same copy number profiles, except for the deletion of a q arm of chromosome 16 and a p arm of chromosome 17 in *in situ* clone 2, suggesting that the *in situ* clone 2 may have derived from ancestral cells of

the *in situ* clone 1 by additional chromosomal deletions. Additional chromosomal amplifications in chromosome 1 (q arm), 6, 7, 8, 10, and X were observed in the invasive clone.

Next, WES was performed for 11 clusters in the three clones to investigate genomic differences at the single nucleotide level (Figure 3.12). The result shows that the tumor cells in the invasive clone had mutations which do not exist in *in situ* clones. Surprisingly, even *in situ* clones had mutations which were not observed in the invasive clone. This may indicate that IDC is derived from an early ancestry of ductal carcinoma *in situ* (DCIS), atypical ductal hyperplasia (ADH), or other benign cells, not directly from DCIS in a linear manner. Moreover, *in situ* clone 2 obtained mutations exclusive to *in situ* clone 1. This supports that *in situ* clone 2 may be derived from the ancestral cells of *in situ* clone 1, as mentioned based on CNA profiles.

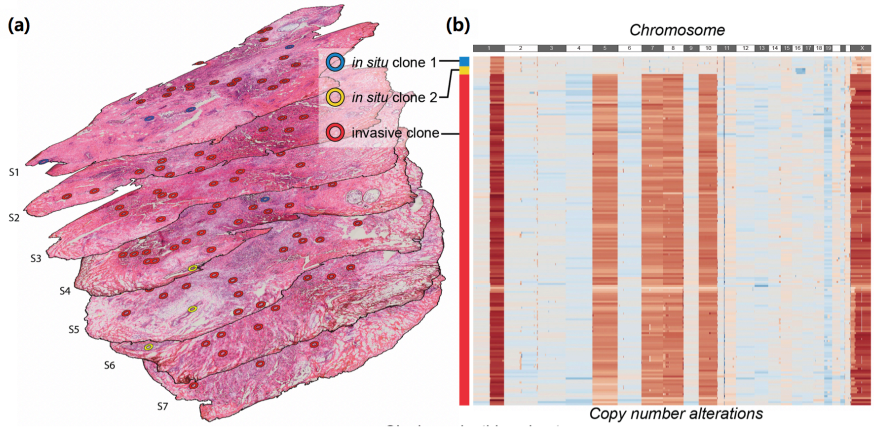

Figure 3.10 The 3-dimensional tumor mass was investigated using PHLI-seq. (a) A total of 177 cell clusters were isolated and sequenced from 7 consecutive tissue sections from the triple-negative breast tumor. (b) Three subclones, which were named as in situ clone 1, 2 and invasive clone, were identified by copy numbers analysis and clustering [25].

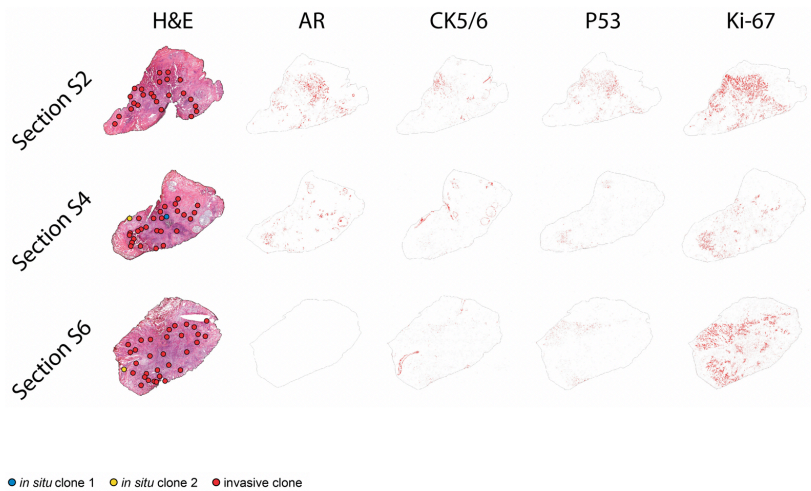

Figure 3.11 Tissue sections were stained by H&E and IHC, before applying PHLI-seq to the triple-negative breast cancer analysis. The serial tissue sections were stained to analyze expression level of AR, CK5/6, Ki-67, and p53 [25].

Within the two *in situ* clones, which were determined by whole genome sequencing, cells with different histopathological characteristics were observed. DCIS and benign usual ductal hyperplasia were included in *in situ* clone 1, while DCIS and ADH were included in *in situ* clone 2. Histopathologic evaluation of the invasive clone showed that every cell clusters in the invasive clone were IDC (Figure 3.13). A three-dimensional reconstruction of the sections allowed identification of *in situ* clone 2 clusters observed in a vicinity in consecutive sections S4, S5, S6, and S7, suggesting that combined analysis of histopathology and spatially resolved genomics-enabled by PHLI-seq has potential to contribute to clinical diagnostics.

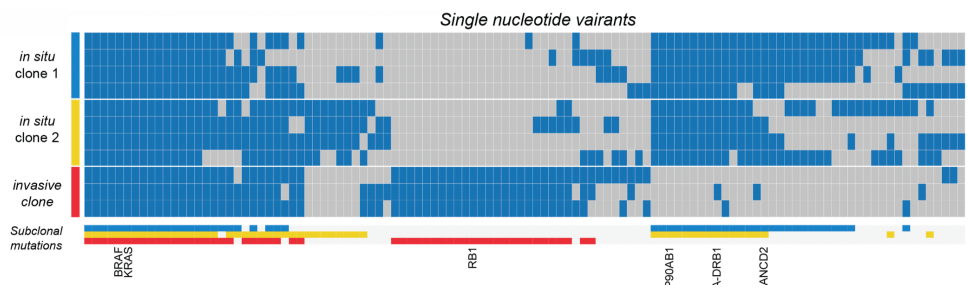

Figure 3.12 Whole-exome sequencing of the selected samples from each subclone. Regardless of in situ or invasive clones, it should be noted that each subclone has its unique somatic SNVs [25].

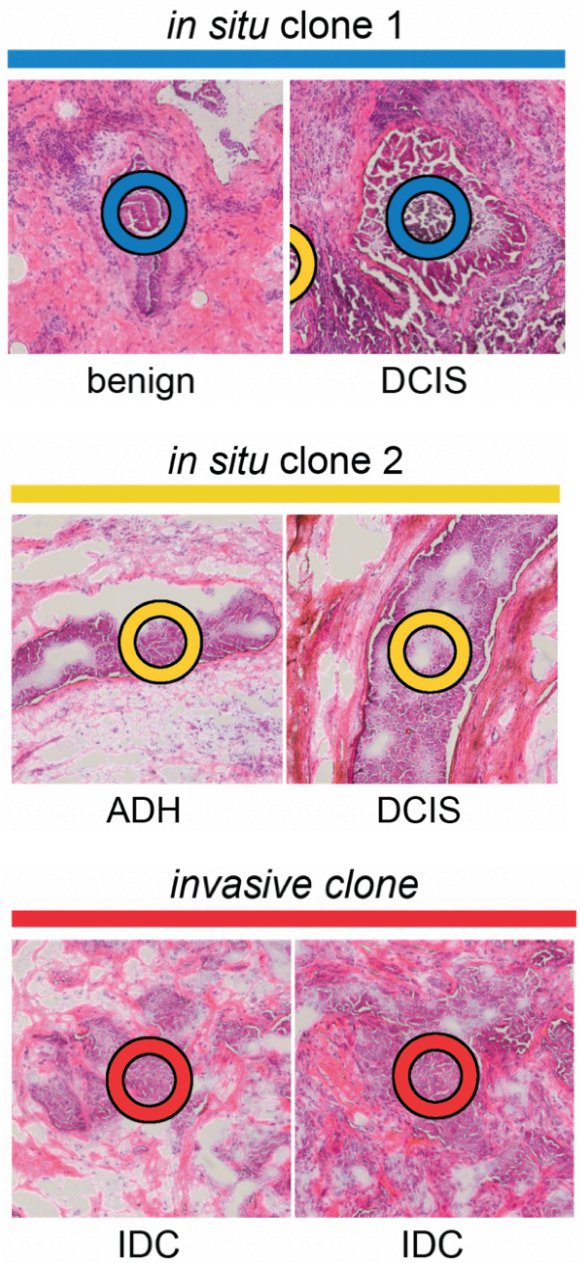

Figure 3.13 In situ clone 1 included DCIS and benign usual ductal hyperplasia, and in situ clone 2 included DCIS and ADH. Tumor cells in Invasive clone were IDC [25].

3.3 A Summary

Using PHLI-seq, three subclones were identified in the HER2-positive breast cancer tissue and discovered subclonal CNAs and point mutations. Here, it was found that the majority of CNAs are shared by all groups, whereas those of SNVs are not. Similar to a prior report [54], this result indicates that most chromosomal rearrangements preceded the generation of point mutations. Moreover, if only CNAs are considered, the development of the tumor seems to follow the model of punctuated evolution with a one-time event and not the model of gradual evolution [16, 18]. However, with regard to SNVs, the evolution cannot be explained by the punctuated evolution model with a one-time event. In total, we visualized the three genetic subclones of HER2-positive breast cancer in the tissue context with their genetic history of the evolution.

PHLI-seq was also applied to the serial breast cancer tissue sections to construct and visualize a cancer genomic map in a three-dimensional spatial context. Three genetic subclones were identified from 177 cell clusters in the 7 mm x 6 mm x 5 mm space, encompassing the histopathologically classified benign cancer cells ADH, DCIS, and IDC. CNA and SNV analyses were carried out by WES to elucidate the evolutionary history of the cancer cells in relation to the histopathological information.

Recently, Navin's group published a research about multi-clonal invasion in breast tumors using topographic single-cell sequencing, which is an excellent example of spatially resolved sequencing [24]. From a methodological point of view, the research group used an UV-based LCM method, and they described that UV cutting parameters influence DNA fragmentation and the transfer failure rate.

In addition, the research group performed somatic CNA analysis from the single cells, but not SNV analysis, which generally requires a much higher quality of amplified DNA. Therefore, it is expected that the advantage of the throughput and IR-based isolation of PHLI-seq will help researchers, who are at the cutting edge of cancer biology, through spatially resolved sequencing.

Chapter 4

Inferring Tumor Evolution of Ovarian Cancer

This chapter describes a case of an ovarian cancer analyzed by PHLI-seq. Epithelial ovarian cancer (EOC) is a silent but mostly lethal gynecologic malignancy [55]. High-grade serous carcinoma is the most common histological EOC subtype. Current treatment strategy for the high-grade serous carcinoma involves a primary debulking surgery followed by chemotherapy to reduce the tumor burden[56], [57]. Recent advances in genomics have revealed the presence of extensive intra-tumor heterogeneity (ITH) in many cancers, including ovarian cancer[58]–[60]. The presence of extensive clonal diversity increases ability of a given tumor to survive upon a strike. Thus, it is thought to be responsible for reduced response to chemotherapy and developing chemoresistance [61]–[63].

Unlike other solid tumors, the primary route of metastasis in EOC patients is the transcoelomic metastasis route, which is a passive process and involves

dissemination of tumor cells from the primary tumor tissue into the peritoneal cavity [64]. Early disseminating clones may exist in the malignant ascites tumor microenvironment (TME) and may form an independent subclonal lineage which can contribute to ITH and metastasis. Both protumorigenic and antitumorigenic factors are known to be enriched in the malignant ascites TME [65]. However, genetic differences between tumor cells in the primary tissue and tumor cells surviving in the ascites TME are not yet fully understood. The sequencing of both the primary tumor in several regions and the associated metastases in ovarian cancer has provided insights into the spatial heterogeneity and shown that metastatic tumors maintain the genetic alterations found in the primary tumor and develop with little accumulation of additional genetic alterations [60]. However, the extent of the genetic heterogeneity within and between the primary tumor and tumor cells found in ascites remains underestimated.

Here, to uncover the genetic heterogeneity of tumor cells in malignant ascites, PHLI-seq technique was applied to individual tumor spheroids which are the common form of tumor cells floating in malignant ascites. Inspired by single-cell analysis, it was hypothesized that genetic profiling of individual tumor spheroids might uncover the heterogeneity within and between the primary tumor and tumor cells in ascites. Individual tumor spheroids were isolated through PHLI-seq technique as described before, and low-depth whole-genome sequencing (WGS) and high-depth whole-exome sequencing (WES) were carried out for ten tumor spheroids and eight primary tumor samples from a high-grade serous (HGS) EOC patient. Similar to the previous cases of breast tumors, somatic copy number alterations (CNAs) and single-nucleotide variants (SNVs) were explored to

determine the tumor evolution and ITH between the primary tissues and the tumor spheroids from the malignant ascites. This study reports the feasibility of analyzing tumor cells in malignant ascites to detect early disseminating EOC clones.

4.1 Sample Preparation

A 42-yr old female patient was diagnosed with primary high-grade serous ovarian cancer (Grade 3, stage IIIC) presented with malignant ascites and peritoneal seeding. For this patient, primary tissues and malignant ascites were collected during primary debulking surgery, then the malignant ascites sample was stored at -80°C. Fresh primary tissues and tumor spheroids were mounted onto ITO-coated glass slides (Figure 4.1). Six primary tissue samples were taken randomly from the solid portions of right ovary and one from left ovary. Blood was collected to serve as the normal control of sequence analysis. Ten tumor spheroids were collected from the malignant ascites and fixed in 10% (v/v) formaldehyde. This study was approved by the Institutional Review Board (IRB) at Seoul National University Hospital (Registration number: 1305-546-487) and performed in compliance with the Helsinki Declaration. Informed consent was obtained from the patient prior to primary debulking surgery to be used in research.

Single tumor spheroids on the slide were isolated by PHLI-seq technique (Figure 4.2a,b). The isolated single tumor spheroids were collected in PCR tubes for further reactions. The isolated single tumor spheroids were lysed by proteinase K and underwent whole genome amplification by MDA method. The amplification was monitored via real-time PCR. The results showed that all the isolated samples yielded successful amplification (10/10). The amplification plot of the non-template control showed that there was no or a negligible amount of carry-over contamination (Figure 4.3). Every reaction yielded over 2 μg of amplified DNA, which was enough to conduct WGS and WES.

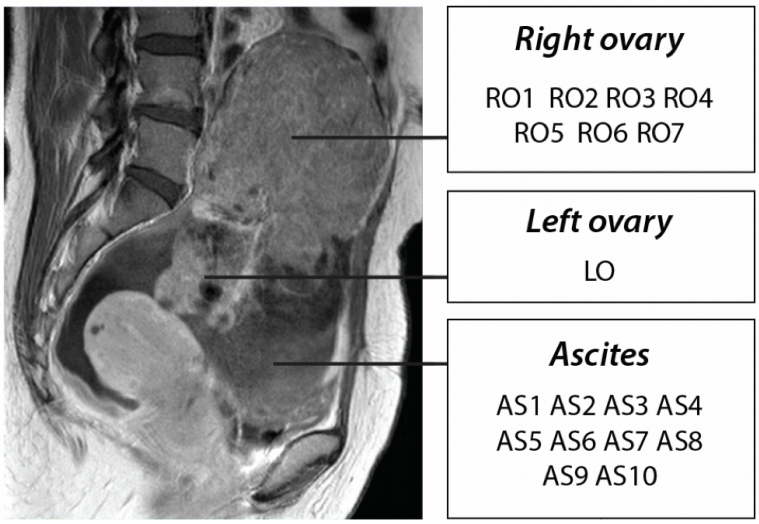

Figure 4.1 A 42-yr old female patient was diagnosed with primary high-grade serous ovarian cancer (Grade 3, stage IIIC) presented with malignant ascites and peritoneal seeding. Seven primary tissue samples at the right ovary (RO1, RO2, RO3, RO4, RO5, RO6, and RO7), one primary tissue sample at the left ovary (LO), and ten tumor spheroids samples at ascites (AS1, AS2, AS3, AS4, AS5, AS6, AS7, AS8, AS9, and AS10) were prepared for sequence analysis [55]

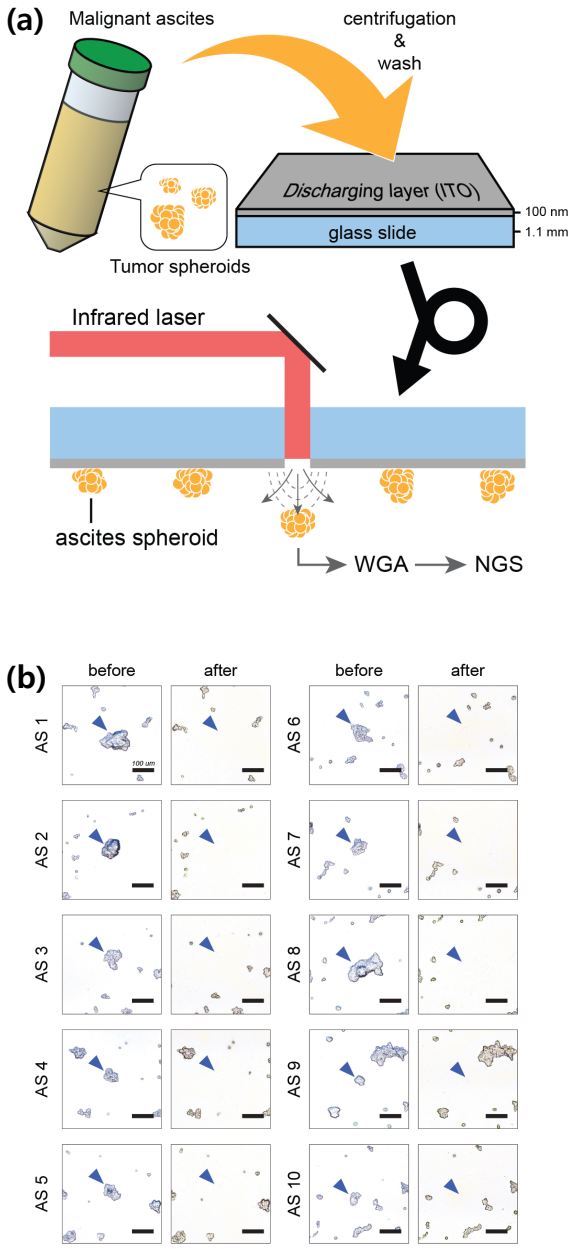

Figure 4.2 The tumor spheroids in the malignant ascites were purified, fixed and prepared on the ITO-coated glass slides. (a) Individual tumor spheroids were isolated and analyzed through PHLI-seq technique. The isolated cells underwent WGA and sequencing. (b) The images before and after isolating tumor spheroids. The scale bars represent 100 um [55]

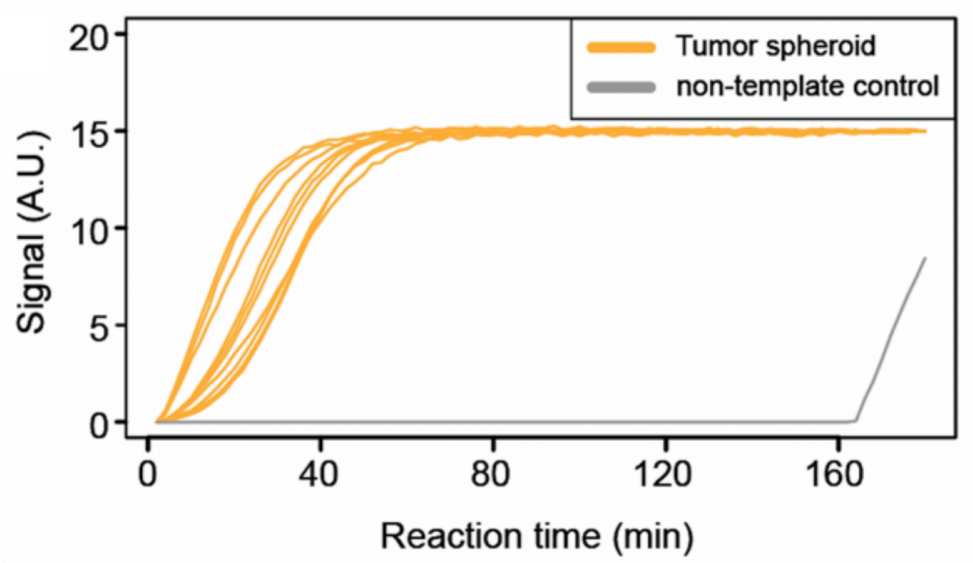

Figure 4.3 Real-time monitoring of whole-genome amplification reactions for the isolated ten tumor spheroids and a non-template control [55].

4.2 Somatic CNA and SNV Analysis

For assessing somatic CNAs of the primary ovarian cancer tissues and the tumor spheroids, low-depth WGS was carried out using the Illumina platform. Through NGS, 8.53 ± 0.879 ($\times 10^6$) reads were produced for each sample, and CNA profiles were generated. Based on the profiles, each sample was clustered by hierarchical clustering (Figure 4.4a). The clustering yielded three distinct genetic subgroups and the primary ovarian cancer tissues (RO 1-7 and LO, named "Primary clone" and colored red) were clustered together. By contrast, two clusters were observed for the tumor spheroids. The samples in one cluster showed a primary-like CNA profile (AS 1-3 and 7-8, named "Ascites clone 1" and colored yellow), but the

other samples presented a normal-like profile (AC 4-6 and 9-10, named “Ascites clone 2”, colored green).

Interestingly, most of the samples shared deletion of FAT1 and amplification of MYC, PARP10, and CYC1 (Figure 4.4b). These genes are reported to be recurrently deleted (FAT1) or amplified (MYC, PARP10, and CYC1) in pan-cancer data [66]. These facts suggest that the shared CNAs might be the driving alterations at the first stage of carcinogenesis. In contrast, KDM5A and NOTCH3, which are known as recurrently amplified genes in ovarian cancer [66], [67], were focally amplified only in the primary clone. These focal amplifications might allow the primary clone to overwhelm the other subclones and finally dominate the left and right ovaries. However, we could not observe a critical focal amplification or a deep deletion exclusive to Ascites clone 1. This implied that other types of alterations might drive Ascites clone 1 to survive or propagate in the peritoneal fluid.

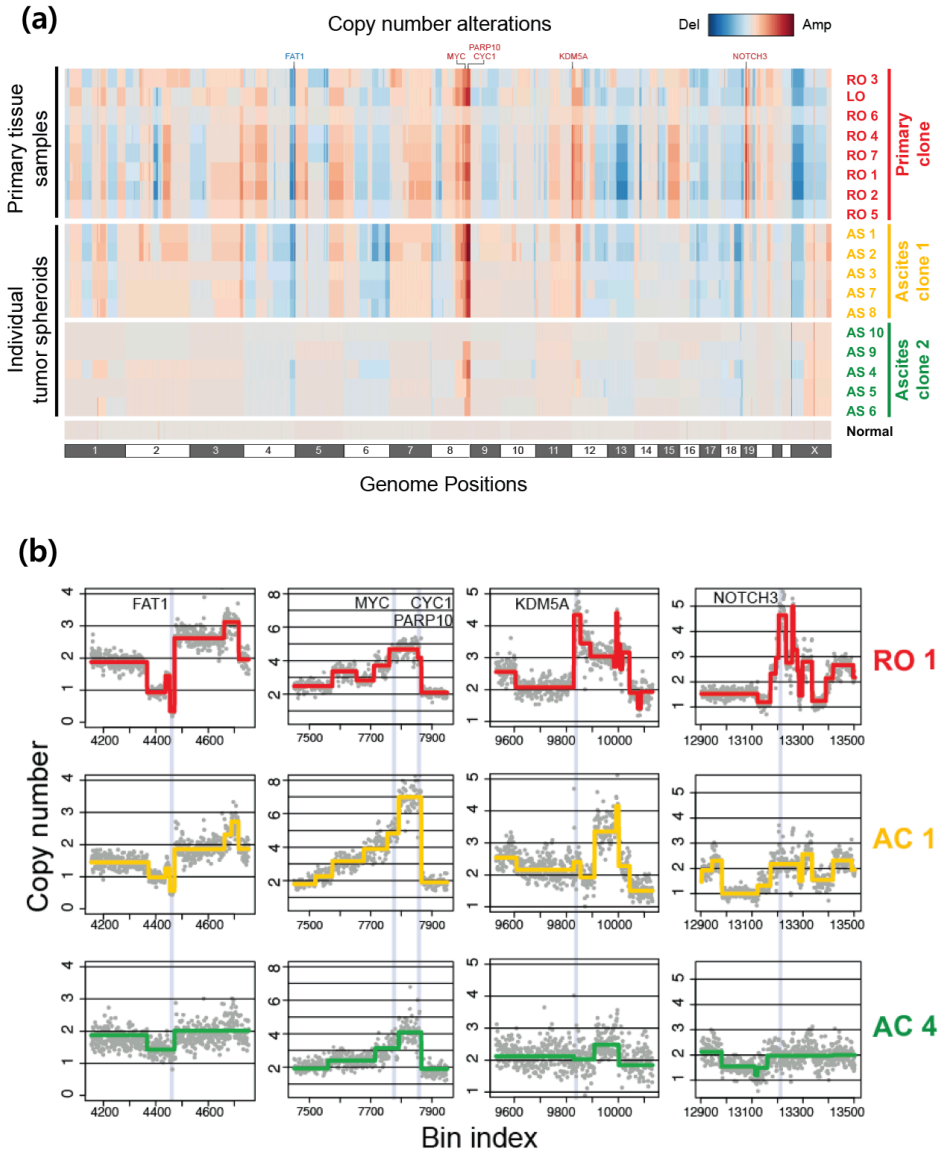

Figure 4.4 CNA analysis based on the genetic subclones of the tumor cells identified via low-depth WGS [55]

Next, somatic SNVs were explored through WES. For each sample, the sequencing run generated 134 ± 21.4 depth of data, covering the whole exome of the human genome. As a result, 171 somatic SNVs were detected by variant calling from all the samples and Figure 4.5a shows the results. Figure 4.5b revealed that 38.6% of the SNVs were common to the primary tumor and tumor spheroids from the ascites, and 35.7% of the SNVs exclusively belonged to primary-only and 25.7% to ascites-only mutations. The exclusive mutations in the Ascites clone suggest that this clone was formed by accumulation of mutations independent of the primary clone. Interestingly, the Ascites clone had a nonsynonymous mutation in the KRAS gene (p.G12D). This single nucleotide substitution leads to an activating KRAS mutation, a known oncogenic mutation associated with the anchorage-independent growth of tumor cells through the acquisition of anoikis resistance to various malignancies [68], [69]. Therefore, the mutation in KRAS in the Ascites clone might provide an additional fitness gain for anchorage-independent survival in the ascites TME. On the other hand, both the Primary and Ascites clones shared somatic SNVs in TP53 and ARID1A, which are well-known driver mutations in ovarian cancer [70], [71]. These mutated genes might be tumor-initiating SNVs at the initial stage of tumorigenesis, in conjunction with the CNAs of FAT1, MYC, PARP10, and CYC1. In addition to these somatic variants, the patient had germline variants in BRCA1 (NM_007294.3:c.1511dupG) and TP53 (NM_001126118:c.C98G), which are well-known susceptibility genes of ovarian cancer and are likely to predispose individuals to ovarian cancer and promote carcinogenesis [72], [73].

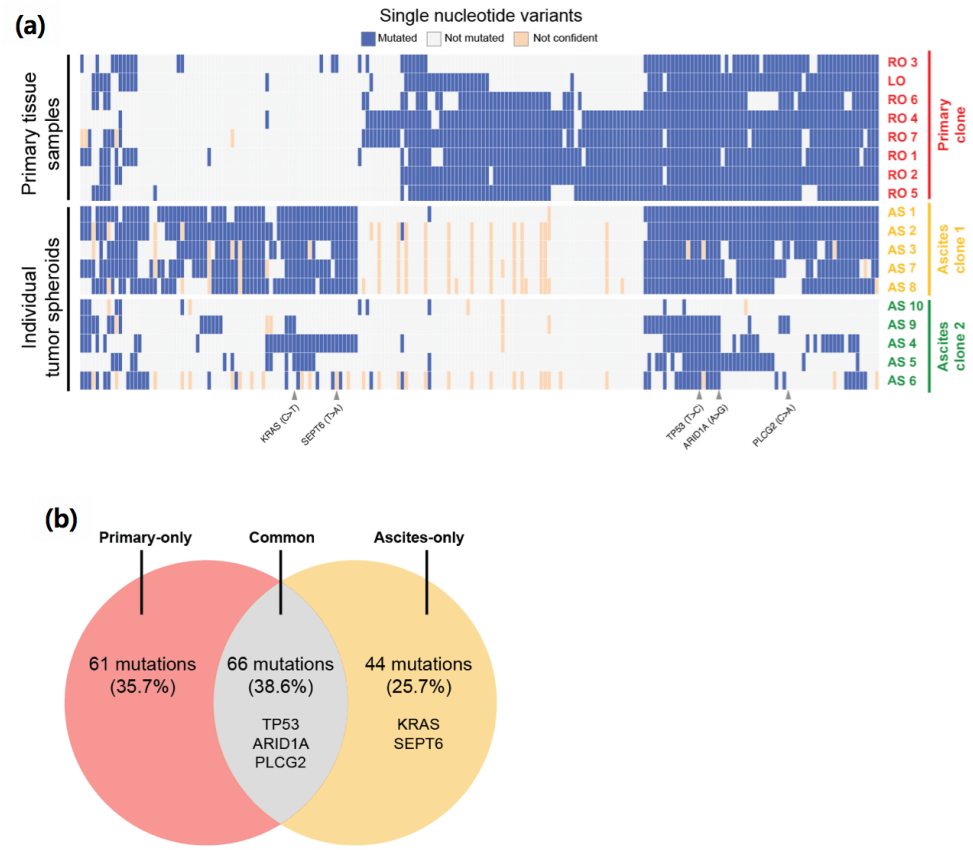

Figure 4.5 Somatic SNVs were discovered using the WES data. The results showed that a significant portion of the SNVs was shared in the Primary clone and Ascites clone 1 [55]

Regarding the CNAs, Ascites clone 2 had no alteration except for amplification of the 8q24 region. Moreover, considering the SNVs, Ascites clone 2 had fewer mutations than the other clusters. Based on these facts, the possibility that normal cells exist in a tumor spheroid was examined. It was assumed that the VAF distribution of Ascites clones 1 and 2 would be similar if the two subclones had a similar proportion of normal cells, while VAFs of Ascites clone 2 would be low if a single tumor spheroid from the clone included a high proportion of normal cells. This idea was tested by plotting the VAF distribution of each sample (Figure 4.6). The results showed that most of the VAF distributions from the Primary clone and Ascites clone 1 were located at a higher range than those from Ascites clone 2. Therefore, it was concluded that the small number of CNAs and SNVs in Ascites clone 2 was not due to their true characteristics but because the proportion of tumor cells was small in the tumor spheroids. Consequently, Ascites clone 2 was excluded from the following phylogenetic analysis.

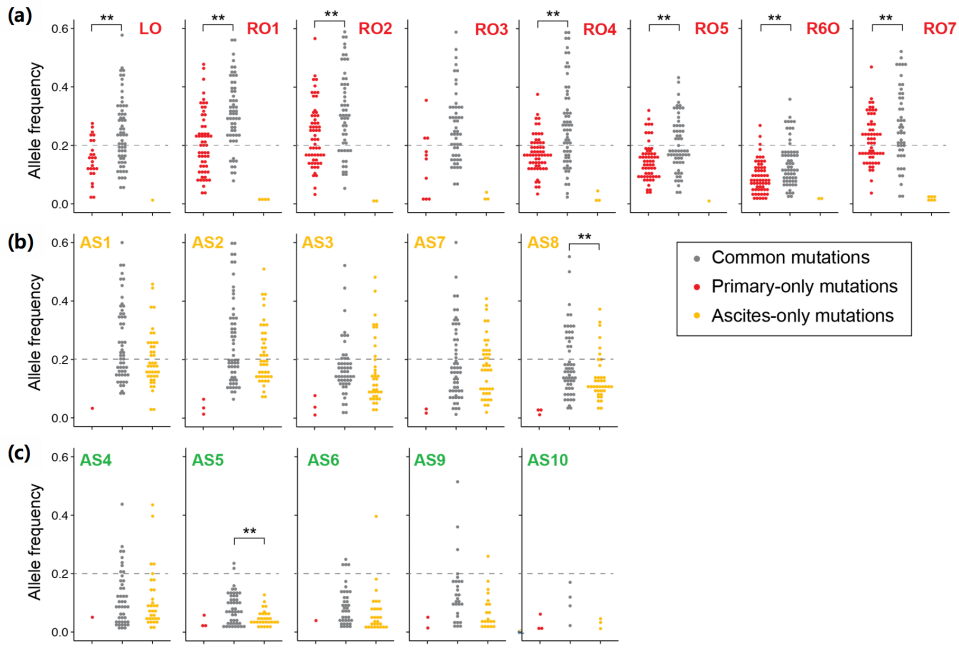

Figure 4.6 The VAF distribution was plotted for each sample of the (a) Primary clones and (b, c) Ascites clones. The mutations were classified into common, primary-only or ascites-only mutations. Common mutations were somatic SNVs which were detected in both the Primary clone and Ascites clones. Primary-only and Ascites-only somatic mutations which were observed only in the Primary clone and Ascites clones, respectively. The results showed that most of the VAF distributions from Ascites clone 2 were located at a much lower range than those from the Primary clone and Ascites clone 1. Therefore, it was concluded that the tumor spheroids in Ascites clone 2 had a large proportion of normal cells in each tumor spheroid [55]

4.3 Inferring Evolutionary Trajectory of the Tumor Cells

Phylogenetic trees were constructed using the CNA and SNV data or the sequenced samples. CNA-based phylogeny was generated as described the other publication [44]. In brief, common chromosomal breakpoints were identified, then a trinary event matrix was constructed. Finally, the maximum parsimony tree was drawn using based on the event matrix (Figure 4.7a). The phylogenetic tree showed that an ancestral cancer clone accumulated CNAs and divided into two subclones, which gained additional exclusive CNAs. Potentially, physically separated and biologically distinct TMEs might drive cancer cells into different alteration statuses.

Although the maximum parsimony tree generated by CNA data provided insights into the process of carcinogenesis, it has a couple of limitations. First, this approach needs to set thresholds to define the amplified, neutral, and deleted status. The phylogenetic tree is significantly affected by thresholds, and there is no golden rule to set the thresholds. Second, the proportion of normal cells in a sample has a substantial impact on the tree because the CNA status might be incorrectly assigned according to the portion of normal cells. For example, the distribution of VAFs of RO6 (Figure 4.4a) shows that RO6 had a large number of normal cells in the sample. In this case, the copy numbers of chromosomes are close to the normal value. Thus, the thresholding led RO6 to be the same as the normal sample, and therefore, we excluded RO6 for constructing the phylogenetic tree based on the CNA data.

Next, a phylogenetic tree was constructed based on the SNV data. This approach does not use manual thresholding, and a phylogenetic tree is less affected by a normal cell portion. Therefore, it was expected that this approach would

provide a more accurate result, compared to the CNA-based approach. The results showed that the cancer cells accumulated mutations as a single clone and divided into two independent clones (Figure 4.7b). Moreover, the phylogenetic tree showed the sequential creation of RO3, LO, and the rest of the Primary clones.

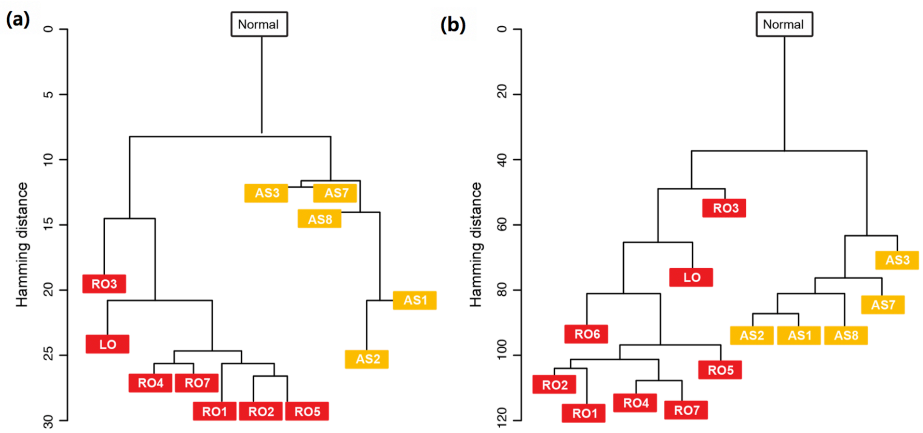

Figure 4.7 Phylogenetic trees were constructed using both the (A) somatic CNAs and (B) somatic SNVs. The two trees presented similar topologies. The result indicated that the Primary clone and Ascites clone 1 were derived from one ancestral clone at the early stage of cancer development [55]

Finally, based on the genomic profiles and inferred phylogenetic tree, the history of the ovarian cancer development and progression was established to understand the tumor evolution and its direction in this patient. As noted before, metastasis of ovarian cancer occurs through a passive process, which initially involves physical shedding of tumor cells from the primary tumor into a peritoneal cavity. Then, the accumulation of ascites facilitates the distant seeding of tumor cells along the peritoneal wall. Given a fixed chance of evolution, two scenarios are possible. One is a monoclonal seeding process and the other is polyclonal seeding. If only certain clones from the primary tumor are fit to survive in the ascites TME, distinct clones, that may have diverged early, may be selected and progress over time in the primary and ascites TMEs, showing a tendency toward independent tumor evolution driven by different TMEs. In contrast, if tumor development is entirely driven by clonal dominance and the physical shedding of tumor cells from the primary tumor occurs by accident, then dominant clones expand in size and others may remain unchanged or become extinct over time at the primary tumor site. With the tumor growth, multiple clones may shed from the primary tumor into ascites. The ascites TME then acts as a reservoir of clonal lineage, and tumor cells in the ascites would represent the entire mutational landscape of a given tumor. For this case, significant genetic differences were observed in the CNAs and SNVs among the primary tissue samples and tumor spheroids. The dominant clones found in the right ovary were absent in the ascites TME. 44 tumor spheroid–specific somatic SNVs were found. Moreover, the comparable allele frequencies between the common mutations and tumor spheroid–specific mutations suggest that the tumor spheroids in the ascites TME

consisted of genetically homogeneous tumor cells compared to the primary tissues. Therefore, it was concluded that the tumor spheroids were from a single subclonal lineage, supporting a mono- and early-seeding origin of the tumor spheroids in this patient. Based on these perspectives, a potential evolutionary trajectory of the tumor was described (Figure 4.8). The initial tumor cells were created at the right ovary to generate the ancestral clone. With further accumulation of mutations, an ancestral clone evolved into two subclones, the first of which was found in the right ovary and metastasized to the left ovary. The second subclone shed into the ascites TME, then became extinct or dominated by the first subclone in the right ovary.

To validate that the Ascites subclone did not exist in the right ovary, four additional primary tissue samples were screened. First, eight different loci of ascites clone specific mutations were selected. Then genotypes of the loci for the four addition primary tissue samples were analyzed (Table 3), and the result showed that the four samples were not matched to Ascites clone. Therefore, it was thought that the genotyping results also support our scenario. In addition, the summary of genome-wide somatic CNAs and SNVs indicated that the tumor cells in the primary tissue and the ascites possessed exclusive alterations as well as common ones. This indicates that the tumor cells in the primary tissue and the ascites were two subclonal lineages, which branched from one ancestral lineage.

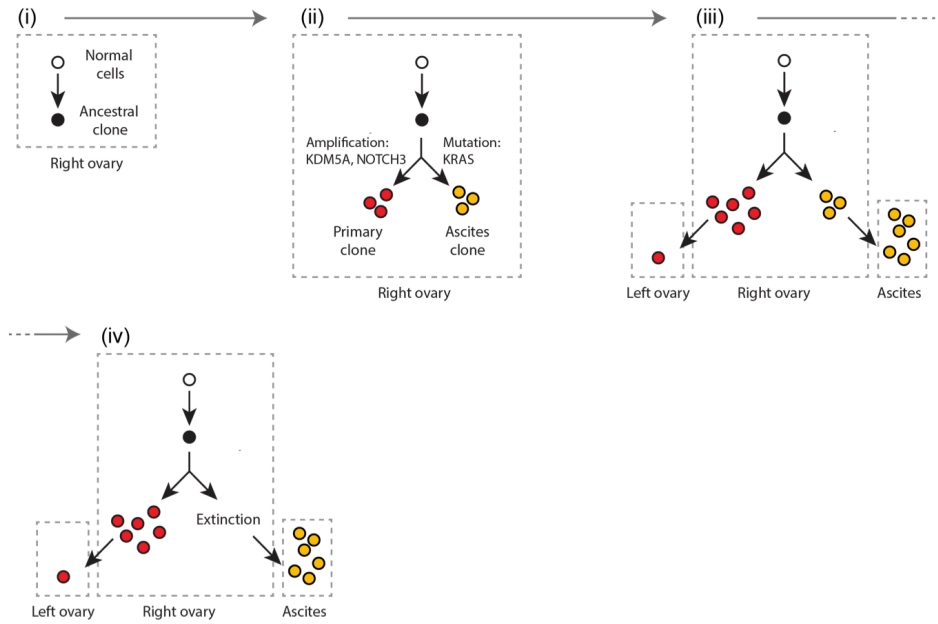

Figure 4.8 The inference of evolutionary trajectory of tumor cells [55].

Table 3 To support that Ascites subclone were not in primary sites, four additional primary tissue samples were sampled and analyzed. Eight different loci where Ascites clone specific mutations were observed were selected. Using the eight loci, each additional primary tissue sample was tested whether it contained Ascites clone specific mutation through genotyping by a capillary sequencing method [55].

locus #	Chr:Pos	expected base change (in Ascites clone)	Genotype at the targeted loci				
			AS 1	Additional primary tissue samples			
				RO8 (Additional)	RO9 (Additional)	RO10 (Additional)	RO11 (Additional)
1	12:27156286	G>T	G/T	G	G	G	G
2	7:100014760	G>A	G/A	G	G	G	G
3	2:48873923	C>G	C/G	C	C	C	C
4	15:27184492	A>C	A/C	A	A	A	A
5	17:54558053	G>C	G/C	G	G	low qual	G
6	17:5404060	G>C	low qual	low qual	low qual	low qual	low qual
7	15:25322219	T>C	T/C	T	T	T	T
8	11:35276357	T>A	T/A	T	T	T	T

4.4 A Summary

In this study, I attempted to determine the presence of genetic heterogeneity within and between the primary tumor and the associated tumor spheroids in the ascites by performing multi-region sequencing of the primary tumor and genetic profiling of the individual tumor spheroids through PHLI-seq. Both WGS and WES were carried out for the primary tumor and tumor spheroid samples. First, ITH levels in eight primary tissues and ten tumor spheroids were discovered through CNA analysis. Also, it was observed that the CNA profiles of the primary and associated tumor spheroids were separated into two distinct genetic clusters, suggesting that the TME may be operative during tumor evolution. Second, it was discovered that a total of 171 somatic SNVs from all the samples, and 66 (38.6 %) of these SNVs were ubiquitous mutations that were common to the primary tumor and tumor spheroids. The rest were either primary-only (61 SNVs, 35.7 %) or ascites-only (44 SNVs, 25.7 %) mutations.

Analysis of VAF distribution indicated the level of normal cell contamination in both the primary tumor and tumor spheroids. Both the Primary clone and Ascites clone 1 showed higher VAF distributions than Ascites clone 2, suggesting that the normal-like CNA and SNV profiles in Ascites clone 2 were due to a high portion of normal cells. These findings could be supported by a paper [74] which suggested the presence of tumor-associated macrophages in the center of tumor spheroids .

Our data showed the presence of genetic heterogeneity within and between the primary tumor and the associated ascites spheroids. Also, our data indicated that the primary and associated ascites spheroids diverged early in tumor

development, and not all the Primary clones disseminated into the ascites TME. However, limitations of the study would be that this study incorporated a single case and provided no insight into distant metastatic sites.

In addition to the importance of genetic differences between tumor cells in primary tissue and those in ascites, knowledge regarding the genetic heterogeneity within the tumor cells in ascites would be valuable. Although genetic diversity of tumor cells in an ascites was not thoroughly studied, the genetic diversity of tumor cells in an ascites may have a large impact on tumor relapse and metastasis, given that transcoelomic spread is the primary route of metastasis in ovarian cancer. However, there has not been attempt to discover the genetic heterogeneity of individual tumor spheroids. In this study, 10 individual tumor spheroids were evaluated, and five of which contained sufficient tumor cells for the analysis. A follow-up study should analyze at least a few tens of individual tumor spheroids per patient to find a clear signature of the genetic heterogeneity in an ascites.

In summary, the genetic plasticity and similarity between a primary tumor and associated tumor spheroids are still not clear, and yet, the nature of the similarity may have profound implications for both tumor progression and therapeutic outcomes in ovarian cancer. Therefore, future prospective studies profiling the genomic information of primary ovarian tumors, distant metastatic tumors, and tumor spheroids to determine the direction of tumor evolution and metastasis of ovarian cancer are warranted.

Chapter 5

Conclusion

In this dissertation, I introduced the new technology that effectively discovers genetic heterogeneity of tumors with accurate mapping of genomic information to the spatial information of the cells. It was shown that PHLI-seq technique outperformed conventional UV laser microdissection techniques in terms of DNA quality and sequencing result of isolated cells and process speed, which could result from wavelength (355 nm) and applied energy (~10 μJ) of the laser used in LPC and LMD. For PHLI-seq, 1064 nm of IR laser was used, and photons in this wavelength have three-times lower energy compared to the UV laser in LPC and LMD. Moreover, ~3 μJ of single laser pulse was used for PHLI-seq, whereas about three-times higher energy per a laser pulse was used for LPC and LMD. Especially in LMD, thousands of laser pulses were irradiated around sample for dissection (80 Hz laser pulse × 30 seconds = 2,400 laser pulses). Moreover, UV-induced damage of laser microdissection, high equipment costs can prevent this technique from being widely adopted by researchers. Although basic optics and mechanical

components are similar in PHLI-seq and laser microdissection, UV optics require special components which can withstand UV damage and correct chromatic aberration, and replacing UV laser optics can reduce manufacturing cost. Therefore, I believe that PHLI-seq is advantageous both in performance and cost.

This new high-throughput technique can precisely bridge the histopathological and genomic alteration landscape by incorporating the genomic alteration map at the single cell level into spatial positions in a tissue, enabling an unprecedented understanding of carcinogenesis. Additionally, by comparing primary cancers with metastatic or recurrent cancers by high-precision analysis of micro-local tissue areas at the single-cell level using PHLI-seq, one can discover novel important molecular features concerning the carcinogenic transition between different tumorigenic stages. As another application, PHLI-seq can provide novel insights at the high-precision single-cell level of interactions between a tumor cell and its microenvironment. These applications have not been clearly elucidated by previous conventional techniques.

Another new innovative improvement supplied by PHLI-seq is its exploitation of molecular identification modality other than H&E, FISH, and IHC, among others. Additionally, PHLI-seq can incorporate machine learning to integrate histopathological information, three-dimensional positions, and genomic alteration landscapes, which will elucidate the relationship between histopathology and genomic and molecular features that have remained obscure based on conventional methods alone. PHLI-seq will make a significant contribution to subclonal evolutionary analyses of carcinogenesis and discovery of novel therapeutic targets at subclonal levels.

Finally, it should be noted that spatial genetic information can affect clinical interpretations, since cancer cells evolve through various geographic conditions and microenvironments in tissues. Depending on geography or microenvironment of tumor cells, the cells can act differently. Thus, the PHLI-seq platform meets the needs of cutting-edge cancer biology, which links histopathology to genomics to enable a synergistic and more precise interpretation of cancer [75], [76]. Although PHLI-seq was applied to genome analysis of tumors in this dissertation, it should be emphasized that PHLI-seq is a general technique for isolating a single-cell or a small number of cells from a sample on a slide in a high-throughput manner. Therefore, most analysis modality could be accompanied by isolation, including transcriptome and epigenome analyses, and I expect that PHLI-seq can serve as an essential tool in a broad range of biological and medical sciences.

Bibliography

[1] S. E. Levy and R. M. Myers, “Advancements in Next-Generation Sequencing,” *Annu. Rev. Genomics Hum. Genet.*, vol. 17, no. 1, pp. 95–115, Aug. 2016.

[2] H. P. Sinn and H. Kreipe, “A brief overview of the WHO classification of breast tumors, 4th edition, focusing on issues and updates from the 3rd edition,” *Breast Care*, vol. 8, no. 2, pp. 149–154, 2013.

[3] F. Markowetz, “A saltationist theory of cancer evolution,” *Nat. Genet.*, vol. 48, no. 10, pp. 1102–1103, 2016.

[4] R. Ke *et al.*, “In situ sequencing for RNA analysis in preserved tissue and cells.,” *Nat. Methods*, vol. 10, no. 9, pp. 857–60, Sep. 2013.

[5] J. P. Junker *et al.*, “Genome-wide RNA Tomography in the Zebrafish Embryo,” *Cell*, vol. 159, no. 3, pp. 662–675, 2014.

[6] K. Achim *et al.*, “High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin,” *Nat. Biotechnol.*, vol. 33, pp. 503–509, 2015.

[7] R. Satija, J. a Farrell, D. Gennert, A. F. Schier, and A. Regev, “Spatial reconstruction of single-cell gene expression data,” *Nat. Biotechnol.*, vol. 33, no. 5, pp. 495–502, 2015.

[8] G. Peng *et al.*, “Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo,” *Dev. Cell*, vol. 36, no. 6, pp. 681–697, 2016.

[9] R. C. Adam *et al.*, “Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice,” *Nature*, vol. 521, p. 366, Mar. 2015.

[10] N. Schaum *et al.*, “Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris,” *Nature*, vol. 562, no. 7727, pp. 367–372, 2018.

[11] H. Liu *et al.*, “Genome-Wide Linkage, Exome Sequencing and Functional Analyses Identify ABCB6 as the Pathogenic Gene of Dyschromatosis Universalis Hereditaria,” *PLoS One*, vol. 9, no. 2, p. e87250, Feb. 2014.

[12] E. Rheinbay *et al.*, “Recurrent and functional regulatory mutations in breast cancer,” *Nature*, vol. 547, p. 55, Jun. 2017.

[13] J. Eberwine, J.-Y. Sul, T. Bartfai, and J. Kim, “The promise of single-cell sequencing,” *Nat. Methods*, vol. 11, p. 25, Dec. 2013.

[14] Y. Wang and N. E. Navin, “Advances and Applications of Single-Cell Sequencing Technologies,” *Mol. Cell*, vol. 58, no. 4, pp. 598–609, 2015.

[15] P. Dalerba *et al.*, “Single-cell dissection of transcriptional heterogeneity in human colon tumors.,” *Nat. Biotechnol.*, vol. 29, no. 12, pp. 1120–7, Dec. 2011.

[16] E. Z. Macosko *et al.*, “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets,” *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[17] G. X. Y. Zheng *et al.*, “Haplotyping germline and cancer genomes with high-throughput linked-read sequencing,” *Nat. Biotechnol.*, vol. 34, p. 303, Feb. 2016.

[18] A. M. Klein *et al.*, “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells,” *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[19] N. Crosetto, M. Bienko, and A. van Oudenaarden, “Spatially resolved transcriptomics and beyond,” *Nat. Rev. Genet.*, vol. 16, no. January, pp. 3–5, 2015.

[20] J. H. Lee *et al.*, “Highly Multiplexed Subcellular RNA Sequencing in Situ,” *Science (80-.).*, vol. 343, no. 6177, p. 1360 LP-1363, Mar. 2014.

[21] R. Ke *et al.*, “In situ sequencing for RNA analysis in preserved tissue and cells,” *Nat. Methods*, vol. 10, no. 9, pp. 857–860, 2013.

[22] I. Gaspar and A. Ephrussi, “Strength in numbers: Quantitative single-molecule RNA detection assays,” *Wiley Interdiscip. Rev. Dev. Biol.*, vol. 4, no. 2, pp. 135–150, 2015.

[23] S. Nichterwitz *et al.*, “Laser capture microscopy coupled with Smart-seq2 for spatial transcriptomic profiling,” *Nat. Commun.*, vol. 7, pp. 1–11, 2016.

[24] A. K. Casasent *et al.*, “Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing,” *Cell*, vol. 172, no. 1–2, p. 205–217.e12, 2018.

[25] S. Kim *et al.*, “PHLI-seq: constructing and visualizing cancer genomic maps in 3D by phenotype-based high-throughput laser-aided isolation and sequencing,” *Genome Biol.*, vol. 19, no. 1, p. 158, 2018.

[26] A. K. Casasent, M. Edgerton, and N. E. Navin, “Genome evolution in ductal carcinoma in situ : invasion of the clones,” *J. Pathol.*, vol. 241, pp. 208–218, 2017.

[27] K. Schütze and G. Lahr, “Identification of expressed genes by laser-mediated manipulation of single cells,” *Nat. Biotechnol.*, vol. 16, no. 8, pp. 737–742, 1998.

[28] V. Espina *et al.*, “Laser-capture microdissection,” *Nat. Protoc.*, vol. 1, no. 2, pp. 586–603, 2006.

[29] M. Vandewoestyne, K. Goossens, C. Burvenich, A. Van Soom, L. Peelman, and D. Deforce, “Laser capture microdissection : Should an ultraviolet or infrared laser be used?,” *Anal. Biochem.*, vol. 439, no. 2, pp. 88–98, 2013.

[30] S. A. López-Haro, M. I. Gutiérrez, A. Vera, and L. Leija, “Modeling the thermo-acoustic effects of thermal-dependent speed of sound and acoustic absorption of biological tissues during focused ultrasound hyperthermia,” *J. Med. Ultrason. (2001)*, vol. 42, no. 4, pp. 489–498, 2015.

[31] C. M. and S. L. and J. L. O. and J. J. G. and L. U. and J. Cárabe, “Microprocessing of ITO and a-Si thin films using ns laser sources,” *J. Micromechanics Microengineering*, vol. 15, no. 6, p. 1271, 2005.

[32] S. L. Jacques, “Optical properties of biological tissues: a review,” *Phys. Med. Biol.*, vol. 58, no. 11, p. R37, 2013.

[33] N. Arneson, S. Hughes, R. Houlston, and S. Done, “Whole-genome amplification by improved primer extension preamplification PCR (I-PEP-PCR),” *Cold Spring Harb. Protoc.*, vol. 3, no. 1, pp. 1–5, 2008.

[34] C. Zong, S. Lu, A. R. Chapman, and X. S. Xie, “Genome-wide detection of single-nucleotide and copy-number variations of a single human cell.,” *Science*, vol. 338, no. 6114, pp. 1622–6, Dec. 2012.

[35] N. E. Navin, “Cancer genomics: one cell at a time,” *Genome Biol.*, vol. 15, no. 8, p. 452, 2014.

[36] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[37] T. Baslan *et al.*, “Genome-wide copy number analysis of single cells,” *Nat. Protoc.*, vol. 7, no. 6, pp. 1024–1041, 2012.

[38] H. Willenbrock and J. Fridlyand, “A comparison study: Applying segmentation to array CGH data for downstream analyses,” *Bioinformatics*, vol. 21, no. 22, pp. 4084–4091, 2005.

[39] H. Shimodaira, “Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling,” *Ann. Stat.*, vol. 32, no. 6, pp. 2616–2641, 2004.

[40] R. Suzuki and H. Shimodaira, “Pvclust : an R package for assessing the uncertainty in hierarchical clustering,” *Bioinformatics*, vol. 22, no. 12, pp. 1540–1542, 2006.

[41] T. C. G and E. Atlas, “Comprehensive molecular portraits of human breast tumours.,” *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[42] C. Kandoth *et al.*, “Mutational landscape and significance across 12 major cancer types.,” *Nature*, vol. 502, no. 7471, pp. 333–9, 2013.

[43] W. Malilas *et al.*, “Cancer upregulated gene 2, a novel oncogene, enhances migration and drug resistance of colon cancer cells via STAT1 activation,” *Int. J. Oncol.*, vol. 43, no. 4, pp. 1111–1116, 2013.

[44] R. Gao *et al.*, “Punctuated copy number evolution and clonal stasis in triple-negative breast cancer,” *Nat. Genet.*, vol. 48, pp. 1119–1130, 2016.

[45] N. Navin *et al.*, “Tumour evolution inferred by single-cell sequencing,” *Nature*, vol. 472, no. 7341, pp. 90–94, 2011.

[46] A. McKenna *et al.*, “The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res.*, vol. 20, pp. 1297–1303, 2010.

[47] M. A. Lodato *et al.*, “Somatic mutation in single human neurons tracks developmental and transcriptional history,” *Science (80-.).*, vol. 350, no. 6256, pp. 94–98, 2015.

[48] D. C. Koboldt *et al.*, “VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing,” *Genome Res.*, vol. 22, no. 3, pp. 568–576, 2012.

[49] K. Cibulskis *et al.*, “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples,” *Nat. Biotechnol.*, vol. 31, no. 3, pp. 213–219, 2013.

[50] T. Ikenoue *et al.*, “Functional Analysis of PIK3CA Gene Mutations in Human Colorectal Cancer,” *Cancer Res.*, vol. 61, no. 11, pp. 4562–4568, 2005.

[51] M. Gymnopoulos, M.-A. Elsliger, and P. K. Vogt, “Rare cancer-specific mutations in PIK3CA show gain of function,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 13, pp. 5569–5574, 2007.

[52] M. W. Schmitt, S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, and L. A. Loeb, “Detection of ultra-rare mutations by next-generation sequencing,” vol. 2012, pp. 1–6, 2012.

[53] S. R. Kennedy *et al.*, “Detecting ultralow-frequency mutations by Duplex Sequencing,” *Nat Protoc*, vol. 9, pp. 2586–2606, 2014.

[54] Y. Wang *et al.*, “Clonal evolution in breast cancer revealed by single nucleus genome sequencing,” *Nature*, vol. 512, no. 7513, pp. 155–160, Jul. 2014.

[55] S. Kim *et al.*, “Evaluating Tumor Evolution via Genomic Profiling of Individual Tumor Spheroids in a Malignant Ascites,” *Sci. Rep.*, vol. 8, no. 1, p. 12724, 2018.

[56] R. Brett M., P. Jennifer B., S. Thomas A., R. Brett M., P. Jennifer B., and S. Thomas A., “Epidemiology of ovarian cancer: a review,” *Cancer Biol. Med.*, vol. 14, no. 1, pp. 9–32, 2017.

[57] P. Jessmon, T. Boulanger, W. Zhou, and P. Patwardhan, “Epidemiology and treatment patterns of epithelial ovarian cancer,” *Expert Rev. Anticancer Ther.*, vol. 17, no. 5, pp. 427–437, May 2017.

[58] L. Ding *et al.*, “Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing,” *Nature*, vol. 481, no. 7382, pp. 506–510, 2012.

[59] M. Gerlinger, A. J. Rowan, S. Horswell, and J. Larkin, “Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing,” *N. Engl. J. Med.*, vol. 366, no. 10, pp. 689–698, 2012.

[60] J.-Y. Lee *et al.*, “Tumor evolution and intratumor heterogeneity of an epithelial ovarian cancer investigated using next-generation sequencing,” *BMC Cancer*, vol. 15, no. 1, p. 85, 2015.

[61] C. Swanton, “Intratumor heterogeneity: Evolution through space and time,” *Cancer Res.*, vol. 72, no. 19, pp. 4875–4882, 2012.

[62] M. A. Bilen *et al.*, “Intratumoral heterogeneity and chemoresistance in nonseminomatous germ cell tumor of the testis,” *Oncotarget*, vol. 7, no. 52, pp. 86280–86289, 2016.

[63] M. Greaves, “Evolutionary determinants of cancer,” *Cancer Discov.*, vol. 5, no. 8, pp. 806–821, 2015.

[64] D. S. Tan, R. Agarwal, and S. B. Kaye, “Mechanisms of transcoelomic metastasis in ovarian cancer,” *Lancet Oncol.*, vol. 7, no. 11, pp. 925–934, 2006.

[65] S. Kim, B. Kim, and Y. S. Song, “Ascites modulates cancer cell behavior, contributing to tumor heterogeneity in ovarian cancer,” *Cancer Sci.*, vol. 107, no. 9, pp. 1173–1178, 2016.

[66] T. I. Zack *et al.*, “Pan-cancer patterns of somatic copy number alteration,” *Nat. Genet.*, vol. 45, no. 10, pp. 1134–1140, 2013.

[67] J. T. Park *et al.*, “Notch3 gene amplification in ovarian cancer,” *Cancer Res.*, vol. 66, no. 12, pp. 6312–6318, 2006.

[68] M. Derouet *et al.*, “Acquisition of anoikis resistance promotes the emergence of oncogenic K-ras mutations in colorectal cancer cells and stimulates their tumorigenicity in vivo.,” *Neoplasia*, vol. 9, no. 7, pp. 536–545, 2007.

[69] M. Rytömaa, L. M. Martins, and J. Downward, “Involvement of FADD and caspase-8 signalling in detachment-induced apoptosis,” *Curr. Biol.*, vol. 9, no. 18, pp. 1043–1046, 1999.

[70] A. A. Ahmed *et al.*, “Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary,” *J. Pathol.*, vol. 221, no. 1, pp. 49–56, 2010.

[71] K. C. Wiegand *et al.*, “ARID1A Mutations in Endometriosis- Associated Ovarian Carcinomas,” *N. Engl. J. Med.*, vol. 363, no. 16, pp. 1532–1543, 2010.

[72] P. L. Welcsh and M. C. King, “BRCA1 and BRCA2 and the genetics of breast and ovarian cancer.,” *Hum. Mol. Genet.*, vol. 10, no. 7, pp. 705–713, 2001.

[73] Y. Xi, C. Liu, and X. Xin, “Association Between a Single Nucleotide Polymorphism in the TP53 Region and Risk of Ovarian Cancer,” *Cell Biochem. Biophys.*, vol. 70, no. 3, pp. 1907–1912, 2014.

[74] M. Yin *et al.*, “Tumor-associated macrophages drive spheroid formation during early transcoelomic metastasis of ovarian cancer,” *J. Clin. Invest.*, vol. 126, no. 11, pp. 4157–4173, 2016.

[75] S. W. Song *et al.*, “One-Step Generation of a Drug-Releasing Hydrogel Microarray-On-A-Chip for Large-Scale Sequential Drug Combination Screening,” *Adv. Sci.*, vol. 0, no. 0, p. 1801380, Nov. 2018.

[76] S. W. Song *et al.*, “Uniform Drug Loading into Prefabricated Microparticles by Freeze-Drying,” *Part. Part. Syst. Charact.*, vol. 34, no. 5, p. 1600427, Mar. 2017.

국문 초록

지금까지는 기술적인 한계로 인해 NGS 를 이용하여 생산된 세포의 유전체 정보가 세포의 조직학적 정보와 연결될 수 없었다. 본 박사학위 논문에서는 단일 세포 혹은 소량의 세포를 조직으로부터 고속으로 분리하여 유전체 분석을 수행하는 PHLI-seq 기술을 소개한다. 그리고 PHLI-seq 기술을 이용하여 암 조직을 대상으로 유전체 정보와 조직학적 정보를 연결하는 연구를 소개한다. 유방암 조직에 대하여 PHLI-seq 을 적용하여 유방암 세포의 유전체 분석을 수행하고 분석된 세포의 유전체 정보와 조직학적 이미지를 연동하여 유방암 조직의 유전적 다양성을 시각화 한다. 또한, 난소암에 대하여 PHLI-seq 을 적용하여 암의 발생 및 진행의 과정을 연구한 내용을 소개한다. 이 연구들을 통하여 암 세포들의 서브클론 형성 및 공간적인 배열이 암 발생 초기에 이루어짐을 알 수 있다.

키워드 : 유전체학, 암, 차세대염기서열분석, 염기서열분석의 공간분해능, 레이저미세절제

학 번 : 2013-30999